Estimation Methodology for Portfolio Construction under Uncertainty

Constantinos Chr. Chappas

Department of Statistical Science

University College London

Thesis submitted for the degree of Doctor of Philosophy

University of London

May 2003

Abstract

Portfolio selection is a financial decision problem faced by all investors. Private investors, companies or financial institutions need to decide on how to invest in assets by selecting a portfolio according to some optimality criterion and under possible constraints. Expressed in mathematical terms, the portfolio optimization problem involves quantities which are usually estimated from historical data. Such estimates are accompanied by uncertainty which, via the optimization process, is transferred to the investment decisions, thus rendering many portfolio estimators unstable or unreliable.

This thesis approaches the problem from two angles. On the one hand, we propose an improvement of the sample moments plug-in estimator through its bootstrap distribution. A robust measure of location of this distribution results, on average, in better out-of-sample performance, especially when the original estimator exhibits high instability, as illustrated by simulations.

On the other hand we propose an alternative way of choosing the optimal intensity of two shrinkage estimators. These estimators stabilize the portfolio by applying shrinkage towards desirable targets. In the first case, these targets are the conventional ones for the mean and the covariance matrix, whereas in the second case we allow for additional market information to be included. Our method again uses bootstrap resamples to account for each estimator's possible out-of-sample performance.

Finally, we consider the problem from a practitioner's perspective by including transaction costs. We exploit a striking similarity between the new optimization problem and the lasso estimator, a variation of the ordinary least squares estimator. We modify accordingly and extend further an existing algorithm for the solution of this problem and present the results. The new algorithm allows for additional constraints on the model coefficients and could be useful in a regression framework when assumptions on the coefficients' sign or magnitude are made.

Acknowledgements

I would like to express my deepest gratitude and appreciation towards my two supervisors, Professor Phil Dawid and Dr Kostas Skouras, both of whom have been extremely helpful during the course of my research. In particular, I am grateful for Kostas' constant guidance, patience and genuine support and Phil's thoughtful comments and invaluable suggestions. They have been my inspiration and I am indebted to both.

Special thanks go to my colleagues and staff at the Department of Statistical Science, University College London, who were always willing to help in any way possible. I will also never forget the belief and continuous encouragement that I received from my close friends in Cyprus and abroad, especially during difficult times. In their own way, they have been a source of strength and perseverance.

Finally, I am eternally grateful to my parents, Christos and Ioanna, for their unconditional love and unwavering support throughout my studies. This journey would not have been a successful one if it had not been for their understanding, their advice and their faith in me. As a small token of my appreciation, this thesis is dedicated to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Portfolio selection tackles the problem faced by all investors: selecting from among a large number of available assets an optimal portfolio and managing it through time. Based on their expectations, news reports, market analysis or simply their sentiments, investors decide on how to invest their money in the stock-market. This decision objective, which is easy to state yet complicated to meet, is the focus of this thesis.

The terms "selecting" and "optimal" indicate that a selection criterion is needed, together with an optimization method. Optimization inputs are also required and considering the fact that there exist "... a large number of available assets", these inputs are bound to be governed by estimation uncertainty, unless investors are blessed with perfect information. Finally, the phrase "... managing it through time" suggests that in a sequential setup additional factors such as current portfolio holdings or transaction costs should be taken into consideration.

The above problem can easily be solved when investors can predict exactly what will happen in the future. The optimization inputs would then be correct, the output would indicate the rational and optimal decision for an investor to take and research such as this would not be necessary. However, optimization inputs are simply calculated guesses; they are estimates. Investors may believe, with some degree of uncertainty, that a particular share price will, for example, increase in the next time period. Since their prediction involves uncertainty, they should account for it when choosing whether and where to invest.

1.1 Objective

The objective of this thesis is to study the current framework on portfolio selection and suggest methods of incorporating uncertainty into the decision process in an effort of reaching robust, but applicable solutions. We view the problem from the perspective of a financial institution. This enforces certain constraints, regarding for example maximum exposure on particular industries, but also lifts some others which are faced by private investors, such as short-selling. In parallel, we acknowledge the necessity of computer code implementing these methods especially in the rapid moving environment of the financial world. Therefore, we develop a programming class of routines which enables the user (an institution, a company or the private investor) to use these procedures for investment purposes.

The thesis is organized in seven chapters. The current chapter introduces the notation used and describes the theoretical problem of portfolio optimization. The second chapter illustrates, through the use of simulations, the problems that arise when the optimization inputs are estimates. It also surveys some current shrinkage methods which attempt to solve this problem. Chapter 3 proposes a general method of improving portfolio estimators by accounting for out-of-sample performance through the use of bootstrap techniques. The fourth chapter proposes a shrinkage portfolio estimator and describes an alternative way of determining the optimal shrinkage intensity. Furthermore, we modify the optimization problem to include market information that may be available to investors and illustrate the new estimation method through a scenario. Chapter 5 focuses on the computational aspect of one-step ahead optimization including constraints and transaction costs. More specifically, it explores the similarity between the constrained optimization objective and the lasso estimator from the statistical literature. An algorithm for the lasso estimator is modified and extended for the problem at hand. Some simulations under different market scenarios are used for comparing the proposed methods with existing estimators in the sixth chapter. Moreover, a case study illustrating sequential optimization is presented. Finally, the seventh chapter concludes.

1.2 Background Information

In this section, we define some of the terms that we will be using in this thesis. Subsequently, we introduce the portfolio optimization problem through a mean-variance objective and a utility framework.

1.2.1 Definitions and Notation

In our notation we have N securities which produce returns on an arbitrary termly basis, for example daily, weekly, monthly or annually. We are solely concerned with relative returns (defined as the ratio of the price change between two consecutive time periods to the original asset price) and hence, we shall be using the term "return" to mean "relative return" or "rate of return" from now on, unless stated otherwise.

We assume that in any particular time period t, the $N \times 1$ vector of returns $\mathbf{x}_t$, consisting of individual securities $(x_{1t}, x_{2t}, \ldots, x_{Nt})^T$ follows a multivariate distribution with unknown mean $\boldsymbol{\mu}_t$ and variance-covariance matrix $\boldsymbol{\Sigma}_t$. Hence the portfolio return at period t, R_t, will depend on the individual securities' returns x_{it} and the proportion of wealth invested in each asset $\mathbf{w}_t = (w_{1t}, w_{2t}, \ldots, w_{Nt})^T$. Often, it is necessary to compare the return of a portfolio with a benchmark. Throughout this thesis we will be using the cash-only benchmark. As a result, the excess portfolio return will also depend on the interest rate r (which we assume to be the same for lending and borrowing). More formally, we can write:

$$R_t = \mathbf{w}_t^T (\mathbf{x}_t - r\mathbf{1})$$

and

$$\mathbf{1}^T \mathbf{w}_t = 1$$

for all t, where $\mathbf{1}$ is a column vector of 1's, with dimensions conforming each time to the rules of linear algebra. Using the same concept of a benchmark, excess wealth at period t is denoted by M_t and is determined by the invested wealth in the previous period, M_{t-1} and the portfolio return R_t through

$$M_t = M_{t-1}(1 + R_t).$$

In general, we can omit the risk-free rate from the above expressions assuming that we adjust the vector of asset returns $\mathbf{x}$ (and therefore its population mean $\boldsymbol{\mu}$) accordingly, by subtracting the interest rate r from each observation. In the finance literature, $\mathbf{x}$ is often termed *excess return.* This leads to simpler algebraic expressions and will be a practice that we will be following from now on.

1.2.2 Mean-Variance Portfolio Theory

The objective of portfolio optimization is to find suitable *portfolio weights* $\mathbf{w}_t$ according to an optimality criterion. Since R_t is a random variable, its moments form a suitable starting point. The expected total return of the portfolio at time t is given by:

$$E\left(\frac{M_t - M_{t-1}}{M_{t-1}}\right) = E\left(R_t\right) = \mathbf{w}_t^T \boldsymbol{\mu}_t,$$

while its variance or *volatility* by

$$Var\left(\frac{M_t - M_{t-1}}{M_{t-1}}\right) = Var\left(R_t\right) = \mathbf{w}_t^T \boldsymbol{\Sigma}_t \mathbf{w}_t.$$

Without any loss of generality we can omit M_{t-1} for now and focus on what is known as the *Mean-Variance Portfolio Theory* (Markowitz, 1952) attributed to Markowitz's idea of mean-variance efficient portfolios. We pause to assume that all risk-averse investors always prefer assets with higher expected return and lower variance. By varying the weights $\mathbf{w}_t$ we obtain portfolios with different combinations of expected return and variance. However, based on the above assumption of mean-variance investors' preferences, not all attainable portfolios are efficient. One can construct two portfolios with the same expected return but different volatility, in which case risk-averse investors would opt for the portfolio with the smallest variance. Alternatively, two portfolios can exhibit the same variance but produce different expected returns; such a scenario would result in all investors choosing the portfolio with the highest expected return.

More formally, the solution to the single-period optimization problem is given by:

$$\mathbf{w}^*(\mu_P) = \arg\min_{\mathbf{w}} \left\{ \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \mid \mathbf{w}^T \boldsymbol{\mu} = \mu_P, \mathbf{1}^T \mathbf{w} = 1 \right\} \tag{1.1}$$

with $\mathbf{1}$ as before, for any chosen expected portfolio return μ_P. Different values of μ_P produce mean-variance efficient portfolios which in turn can be plotted on the *expected return – standard deviation* plane and trace out the efficient frontier, shown in Figure 1.1.

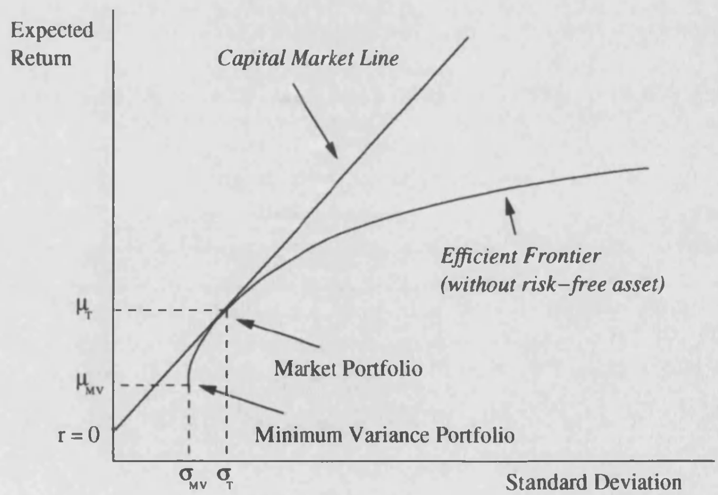

Figure 1.1: Portfolio Expected Return and Standard Deviation

The efficient frontier is the locus of those portfolios that are fully invested and satisfy objective (1.1). Investors may choose to invest part of their wealth on the risk-free asset, represented in Figure 1.1 by the point $(0,0)$ since we are dealing with excess returns. The portfolio optimization problem now is somewhat modified: the existence of the risk-free asset necessitates the selection of a portfolio on the straight line passing through the point $(0,0)$ and is tangent on the original efficient frontier. This straight line is called *capital market line* and consists of portfolios which have the same relative allocation of risky assets but a different proportion of wealth invested in the risk-free asset. The tangency point is called *tangency* or *market portfolio.*

An alternative way of viewing the tangency portfolio is through one important property. It maximizes what is known as *Sharpe's Ratio* (initially introduced in Sharpe (1966) as *reward-to-variability ratio*) of expected portfolio return (in excess of the risk-less return) to its standard deviation, i.e.

$$\mathbf{w}_T = \arg\max_{\mathbf{w}} \left\{ \frac{\mathbf{w}^T \boldsymbol{\mu}}{(\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w})^{\frac{1}{2}}} \right\}$$

subject to the full investment constraint, $\mathbf{1}^T\mathbf{w} = 1$. In Figure 1.1, this is represented by the slope of the line joining the origin and any portfolio on the efficient frontier. It is obviously maximized at the tangency portfolio. It can be shown (see Appendix A.1) that

$$\mathbf{w}_T = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}. \tag{1.2}$$

Note that maximizing the Sharpe ratio is consistent with our, so far, assumptions of mean-variance preferences: the Sharpe Ratio is an increasing function of expected return and a decreasing function of portfolio risk.

An interesting byproduct of Sharpe Ratio optimization occurs when we restrict ourselves to Normally distributed asset returns $\mathbf{x} \sim \mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We can write:

$$\arg\max_{\mathbf{w}} \left\{ \frac{\mathbf{w}^T \boldsymbol{\mu}}{(\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w})^{\frac{1}{2}}} \right\} = \arg\max_{\mathbf{w}} \{\Pr(R > 0)\}$$

where $R = \mathbf{w}^T\mathbf{x} \sim \mathcal{N}\left(\mathbf{w}^T\boldsymbol{\mu}, \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}\right)$, since

$$\Pr(R > 0) = \Pr\left(\frac{R - \mathbf{w}^T\boldsymbol{\mu}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{1}{2}}} > -\frac{\mathbf{w}^T\boldsymbol{\mu}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{1}{2}}}\right) =$$

$$\Pr\left(z > -\frac{\mathbf{w}^T\boldsymbol{\mu}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{1}{2}}}\right) = \Phi\left(\frac{\mathbf{w}^T\boldsymbol{\mu}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{1}{2}}}\right)$$

where $z \sim \mathcal{N}(0, 1)$. Since the cumulative density function $\Phi(z)$ is a monotonically increasing function of z, then maximizing $\Phi(z)$ is equivalent to maximizing z. In this case, z is a portfolio's Sharpe Ratio while $\Phi(z)$ is equal to the probability of obtaining a positive portfolio return (under Normality assumptions). This once again illustrates the competing nature between portfolio risk and return.

In the absence of the risk-free asset (not to be confused with the so far consideration of excess returns), or under an alternative optimality criterion, investors may opt for that portfolio which is subject to the overall lowest risk i.e.

$$\mathbf{w}_{GMV} = \arg\min_{\mathbf{w}}\left\{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}\right\}$$

This is known as the *global minimum variance portfolio*, and its solution is given by

$$\mathbf{w}_{GMV} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}. \tag{1.3}$$

Note that now $\mathbf{w}_{GMV}$ does not depend on $\boldsymbol{\mu}$, since the investor is only interested in the volatility of the portfolio.

1.2.3 Utility Theory

Similar results can be obtained through a utility framework. In fact, this approach allows for a more general analysis and can accommodate particular examples of interest such as the portfolio weights of equations (1.2) and (1.3). We now assume that investors have their own utility curves which depend on the utility they derive from an uncertain monetary reward. We are interested in the behaviour of risk-averse investors and therefore we shall only be considering concave utility curves. Each family of curves is described by parameters which generally relate to individual investors' attitude towards risk. Rational investors choose optimal portfolios according to the expected utility that they will derive from each possible portfolio.

One general class of utility functions (see, for example, Ingersoll (1987)) that are of interest is defined as:

$$U(R_t) = \frac{1-\gamma}{\gamma}\left(\frac{\kappa R_t}{1-\gamma} + b\right)^{\gamma}, \; b > 0, \tag{1.4}$$

for suitable choices of κ, b and γ. This class of utility functions gives rise to a number of families describing investor preferences. For example, a popular choice of a utility function is the *negative exponential utility* family, obtained by setting $b = 1$, $\gamma \to -\infty$ and using the fact that $\lim_{k\to\infty}\left(1 + \frac{z}{k}\right)^k = e^z$. Equation (1.4) then reduces to $-e^{-\kappa R_t}$ but without any loss of generality we can shift the utility upwards to make it strictly positive. Hence, the negative exponential utility is defined as

$$U(R_t) = 1 - e^{-\kappa R_t}, \; \kappa > 0$$

where κ is the coefficient of risk aversion and R_t is the portfolio return, as before. We can show that, if R_t is Normally distributed, an investor who maximizes the expected utility of the total portfolio return will choose weights $\mathbf{w}_U$ such that:

$$\mathbf{w}_U = \arg\max_{\mathbf{w}} \left\{ E\left[U(R_t)\right] \mid \mathbf{1}^T\mathbf{w} = 1 \right\}$$

which results in

$$\mathbf{w}_U = \arg\max_{\mathbf{w}} \left\{ \mathbf{w}^T\boldsymbol{\mu} - \frac{\kappa}{2}\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \mid \mathbf{1}^T\mathbf{w} = 1 \right\}. \tag{1.5}$$

Objective (1.5) is equivalent to objective (1.1) for suitable κ and μ_P. Furthermore, we can show that objective (1.5) leads to the same optimal portfolio $\mathbf{w}_U = \mathbf{w}_T$ described by equation (1.2) for a specific choice of the coefficient of risk aversion $\left(\kappa = \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$. On the other hand, as $\kappa \to \infty$ the optimal $\mathbf{w}_U \to \mathbf{w}_{GMV}$. As a result Sharpe Ratio optimization and portfolio variance minimization are both consistent with expected utility optimization, but using the utility framework enables us to incorporate additional features to the optimization.

The general class of utilities in equation (1.4) includes additional utility families. Other examples include the *quadratic utility* family (obtained by setting $b = 1$ and $\gamma = 2$) while the *power utility* family is obtained for values of $b = 0$ and $\gamma < 1$. A special case of the power utility family is the *logarithmic utility function* at $\gamma = 0$.

1.3 Possible Criticisms

Mean-variance portfolio theory has received criticism regarding certain issues. One important criticism concerns the use of the portfolio's standard deviation as a measure of risk. An investor's utility should be a decreasing function of the portfolio's standard deviation, thus reflecting the investor's preferences regarding deviations from the mean return. However, although large negative returns are undesirable, large positive returns should be considered as pleasant surprises. Accordingly, one usual criticism levelled at this approach is for a more suitable measure of risk to be used. Such risk measures include the semi-variance which is computed as the sum of the squared deviations *below* the mean divided by the total number of observations, or the value-at-risk (VaR) which is the amount of money that can be potentially lost over a specified period of time with a pre-determined probability.

Markowitz (1991) discusses the pros and cons of using the square root of the semi-variance rather than the usual standard deviation as a measure of portfolio risk. One important aspect of portfolio selection, which is often overlooked when this criticism is levelled, is the possibility of short sales or negative portfolio weights[1]. It can be argued that since our analysis will

[1] A *short sale* is effectively the sale of a stock that the investor does not own (on the "promise" that the investor will buy it back at a later stage). Investors sell shares short when they believe that their respective prices will fall. Such practice is usually not

allow for these, measures of downside risk will not be necessarily better. A large positive return in a particular asset will be catastrophic for an investor who expects the particular asset to fall and consequently has sold the asset short. Therefore, we will consider the standard deviation to be more suitable for our purposes.

Another source of criticism is the fact that the mean-variance efficient frontier framework allows only for single-period optimization. Therefore, investors act based on what the optimal decision *for the next period* is. This may not necessarily be the rational decision to take if, for example, the next k periods were taken into consideration.

Multi-period utility functions may be used to solve this problem although these are beyond the scope of this research. Nevertheless, this thesis takes one intermediate step in sequential decision making by investigating the effect of transaction costs on portfolio selection through the utility framework, in a later chapter.

performed by individuals but by institutions. For our purposes, we represent a short sale by a negative element in the portfolio weights vector $\mathbf{w}$.

Chapter 2

Portfolio Estimation

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown to investors and therefore need to be estimated. Having observed a sample of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, the population moments are usually estimated by their sample counterparts $\bar{\mathbf{x}}$ and $\mathbf{S}$ respectively with

$$\bar{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t, \quad \mathbf{S} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^T.$$

Using only the sample moments to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ implies that the sample provides all information needed to predict the behaviour of the assets' returns. An implicit assumption that is also made is that the ordering of the data is not important. In other words the sample should not be considered as a time-series. This is justifiable when the sample relates to the short term and hence expected return is constant through that short space of time.

Although these two assumptions may provide grounds for criticism, our analysis focuses on ways of using the estimates for the expected return and covariance matrix for portfolio construction rather than finding appropriate models to predict the asset returns. If better estimates become available which depend, for example, not only on historical prices but also on other factors such as volume, market sector or firm characteristics, a similar methodology can be employed.

Hence, we proceed using $\bar{\mathbf{x}}$ and $\mathbf{S}$ as estimates of the population moments. In this section we shall be investigating the usefulness of the sample moments estimator of the tangency portfolio and offering some insight on the reasons of its failure.

2.1 The Sample Moments Estimator

The obvious way of estimating the tangency portfolio of equation (1.2) is through the sample moments estimator:

$$\hat{\mathbf{w}}_{SM} = \frac{\mathbf{S}^{-1}\bar{\mathbf{x}}}{\mathbf{1}^T\mathbf{S}^{-1}\bar{\mathbf{x}}}. \tag{2.1}$$

This estimator maximises the in-sample Sharpe Ratio but, as noted by Michaud (1998), is accompanied by a large degree of variability[1]. One way of exploring the instability of this estimator (and consequently, its out-of-sample performance) is through Monte Carlo simulations.

Let us suppose that we have observed a set of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. We create B simulated samples, each of size T from a $\mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. For each simulated sample i we estimate the mean $\bar{\mathbf{x}}^{(i)}$, and covariance matrix $\mathbf{S}^{(i)}$ before calculating the sample moments estimator $\hat{\mathbf{w}}_{SM}^{(i)}$ given by equation (2.1).

The set of estimators $\left\{\hat{\mathbf{w}}_{SM}^{(i)}\right\}$ for $i = 1, \ldots, B$ can be used to estimate the variance of the estimator $\hat{\mathbf{w}}_{SM}$ and any other properties of its sampling distribution. In the following numerical example, we illustrate the behaviour of the sample moments estimator and compare its empirical distribution with the Normal distribution. There are 4 assets available to the investor. For each simulation $T = 120$ "monthly" data points were simulated from a $\mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \frac{1}{12}(0.08, 0.06, 0.04, -0.01)^T$$

and

$$\boldsymbol{\Sigma} = \frac{1}{12}\operatorname{diag}\left(0.24^2, 0.18^2, 0.16^2, 0.15^2\right)$$

[1]In fact, Michaud (1998) working with sign-constrained portfolios, starts by ordering the efficient frontier portfolios in terms of their expected return. He then argues that portfolios around the middle of the efficient frontier will exhibit a high degree of variability, and this is the area where the sample tangency portfolio will lie in.

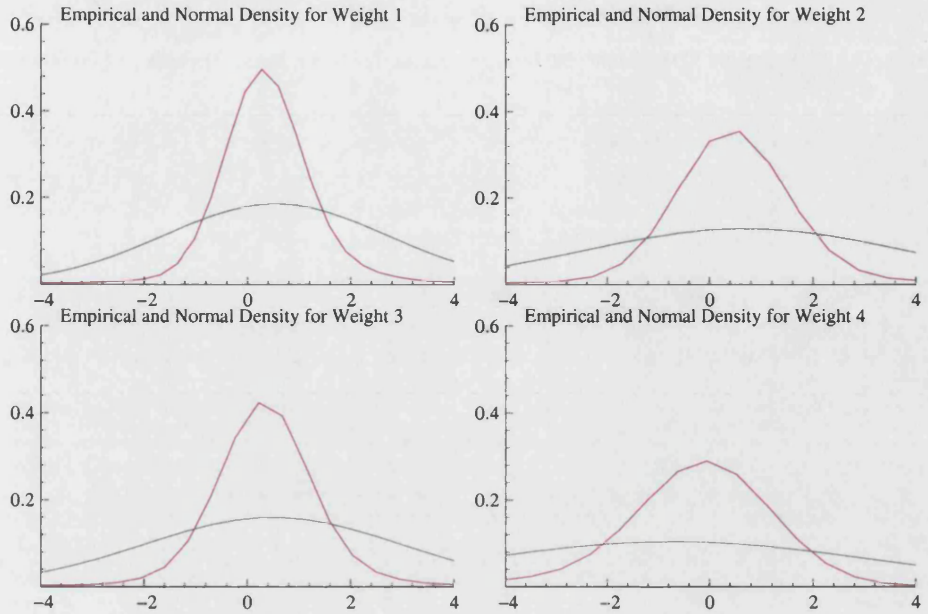

Figure 2.1: Empirical Distribution of the Sample Moments Estimator

where diag $(\mathbf{y})$ denotes a square matrix with the elements of vector $\mathbf{y}$ on its main diagonal. So, for example, the first asset's annualised return would be 8% with a 24% standard deviation. Based on $B = 500$ simulations, the empirical marginal densities for each of the elements in $\hat{\mathbf{w}}_{SM}$ are shown in the four panels of Figure 2.1 by the red curves.

We firstly note the wide range of possible values that the elements in $\hat{\mathbf{w}}_{SM}$ may take. This is augmented by the fact that the are no restrictions on their sign or magnitude (apart from the full investment constraint). As a result, not much information can be extracted by any point estimate $\hat{\mathbf{w}}_{SM}^{(i)}$ about the location of the "optimal" estimator $\hat{\mathbf{w}}_{SM}$.

The shape of the empirical distribution provides an alternative point of interest. In each of the panels, the green curve plots the density of a Normal distribution with the same mean and variance as each of the marginal empirical distributions. In terms of skewness, the empirical densities seem to be symmetric but in terms of kurtosis they tend to exhibit heavy tails. This would point towards an unstable mean as it would be affected by extreme long or short positions in the $\hat{\mathbf{w}}_{SM}^{(i)}$ vectors. It is precisely this feature for which portfolio optimization is notorious.

We can examine the same problem from an additional viewpoint. Figure 2.2 plots the estimated $\hat{\mathbf{w}}_{SM}^{(i)}$ portfolios[2] under the "true" Markowitz efficient frontier generated by the simulation parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The "true"

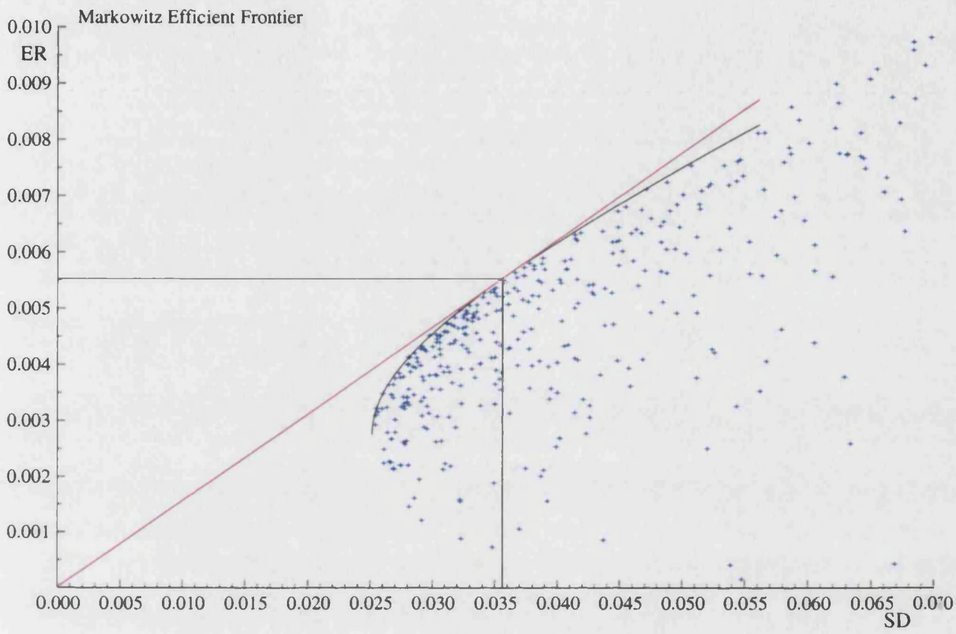

Figure 2.2: Tangency Portfolio Variability under the Efficient Frontier

market portfolio lies at the intersection of the vertical and horizontal lines, at the tangency point on the Markowitz efficient frontier. However, using the sample moments estimator, based on each of the simulated samples, we produce portfolios which are vastly different than the optimal one. Their differences are not only in terms of their composition (i.e. the portfolio weights) but also in terms of their expected return and risk. Furthermore, these portfolios quite often lie well within the frontier and are therefore inefficient.

Similarly, Figure 2.3 plots the respective sample global minimum variance portfolios for comparison. For each simulated sample, the estimated portfo-

[2]Figure 2.2 only shows the part of the efficient frontier around the sample tangency portfolio. In fact, some of the estimated portfolios lay well outside the range of expected return - standard deviation values shown here, but were omitted for clarity.

lios are those which result in the overall lowest risk, derived from the sample equivalent of equation (1.3). Most simulated samples lead to portfolios which, under the true efficient frontier, lie close to the theoretical minimum variance portfolio. It is evident that the instability of the sample moments estimator for the tangency portfolio is even more striking when compared to the situation depicted by Figure 2.3. The minimum variance portfolio exhibits much less variability under the efficient frontier which will prove to be helpful at a later stage.

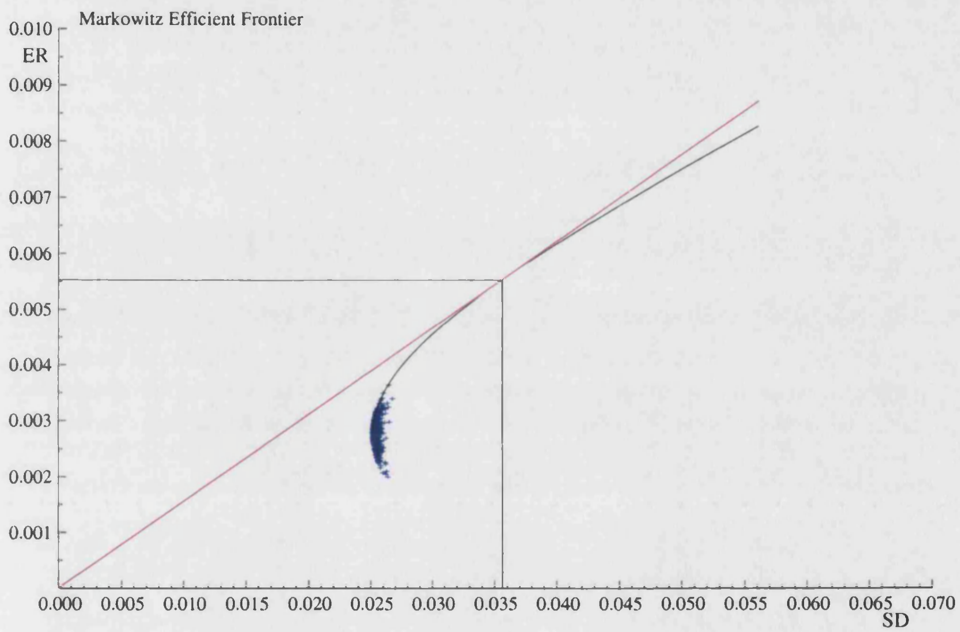

Figure 2.3: GMV Portfolio Variability under the Efficient Frontier

The above simulation study highlights the instability which accompanies the sample tangency portfolio, $\hat{\mathbf{w}}_{SM}$. Although not conclusive, it indicates potential problems when using this investment strategy. Furthermore, it suggests the possible benefits of exploiting the stability of the sample minimum variance portfolio.

2.2 Taylor's Approximation

The problem of instability in the sample moments estimator can be investigated via a multivariate Taylor Series expansion. Let us assume that $\mathbf{\Sigma}$ is known. Then $\hat{\mathbf{w}}_{SM}$ is given by:

$$\hat{\mathbf{w}}_{SM} = \frac{\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}}.$$

Based on the assumption that $\mathbf{x} \sim \mathcal{MN}(\boldsymbol{\mu}, \mathbf{\Sigma})$, each element of $\hat{\mathbf{w}}_{SM}$ is now the ratio of two correlated Normal variables since:

$$\mathbf{\Sigma}^{-1}\bar{\mathbf{x}} \sim \mathcal{MN}\left(\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \frac{1}{T}\mathbf{\Sigma}^{-1}\right),$$

$$\mathbf{1}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}} \sim \mathcal{N}\left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \frac{1}{T}\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right)$$

and

$$Cov\left(\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}, \mathbf{1}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}\right) = \frac{1}{T}\mathbf{\Sigma}^{-1}\mathbf{1}.$$

We can show (see Appendix B.1) that the second order Taylor's approximation around the mean leads to the following expression for the expectation of the ratio estimator,

$$\begin{aligned} E\left(\hat{\mathbf{w}}_{SM}\right) &\approx \frac{\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}} + \frac{1}{T}\left(\frac{\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}} - \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}\right) \times \frac{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2} \\ &= \mathbf{w}_T + \frac{c}{T}\left(\mathbf{w}_T - \mathbf{w}_{GMV}\right). \end{aligned}$$

where $c = \left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right)/(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2$. The first order expression for its variance is given by:

$$Var\left(\hat{\mathbf{w}}_{SM}\right) \approx \frac{1}{T}\mathbf{\Sigma}^{-1}\left(\mathbf{I} + \mathbf{K}\right) \times \frac{1}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2},$$

where $\mathbf{I}$ is the identity matrix and

$$\mathbf{K} = \left(\frac{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}\right)\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1} - \frac{\boldsymbol{\mu}\,\mathbf{1}^T\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-1}\mathbf{1}\,\boldsymbol{\mu}^T}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}.$$

The most interesting issue which arises from this investigation is exactly the one concerning the instability of $\hat{\mathbf{w}}_{SM}$. First of all, the estimator is asymptotically unbiased as

$$\lim_{T\to\infty} E\left(\hat{\mathbf{w}}_{SM}\right) = \frac{\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}.$$

The second term in the approximation for $E(\hat{\mathbf{w}}_{SM})$ is an adjustment of each of the vector elements which directly depends on the difference between the theoretical tangency and global minimum variance portfolios. If these are very similar, something which translates to $\boldsymbol{\mu}$ being approximately proportionate to $\mathbf{1}$ (i.e. the available assets have about the same relative return), then the effect of the second term will be small. Moreover, the second term is inversely proportional to the ratio $(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})/(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})$ which is the expected return of the global minimum variance portfolio. Once again, if this is large (i.e. a large return can be expected at the lowest possible risk) then the second term would again lack effect. In general, the problem is directly related to the unknown population quantity $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. When this is small (and its square is even smaller) the effect of the second term in the expectation is large, and a very large sample size T is required to preserve stability.

A similar conclusion regarding the effect of the quantity $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ can be reached for the estimator's approximate variability. The large variability will render the sample moments estimator for the tangency portfolio, unsuitable. Finally one should note that, in real life, an additional source of uncertainty will be the unknown true covariance matrix $\boldsymbol{\Sigma}$, which was assumed known for the illustrative objective of this analysis.

It is important here to mention what is often called as the *mean blur* (Luenberger, 1997). The sample size T may be "increased" by using a higher frequency of the available historical data. For example, one might suggest using weekly instead of monthly data for better estimation. This will not affect the quantity $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ since it is dimensionless. However, estimation of the population mean becomes harder as the data frequency increases. This is evident in the univariate scenario where the coefficient of variation (defined as the ratio of standard deviation to the sample mean) is an increasing function of the data frequency. Since the coefficient of variation is a measure of dispersion relative to the mean, using higher frequency data, in fact, worsens the estimation problem.

2.3 Current Solutions

Having exposed the problems related to portfolio optimization, we now turn our attention to some methods, currently in use to overcome these problems. The general idea is to improve estimation by enforcing some structure on the optimization inputs. This can be achieved by either imposing certain constraints, or by applying shrinkage to the original optimization inputs towards more stable targets.[3]

2.3.1 Constrained Optimization

Before surveying the shrinkage estimation literature, we consider the case for constrained optimization. The solution to objective (1.5) can be stabilised by enforcing certain constraints on the nature of the portfolio weights. A usual one is the non-negativity constraint which prohibits short sales. In this case the optimization problem becomes:

$$\mathbf{w}_{QP} = \arg\max_{\mathbf{w}} \left\{ \mathbf{w}^T \boldsymbol{\mu} - \frac{\kappa}{2} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \mid \mathbf{1}^T \mathbf{w} = 1,\ \mathbf{w} \geq \mathbf{0} \right\}$$

$$\mathbf{w}_{QP} = \arg\max_{\mathbf{w}} \left\{ \mathbf{w}^T \boldsymbol{\mu} - \frac{\kappa}{2} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \mid \mathbf{1}^T \mathbf{w} = 1,\ \mathbf{w} \geq \mathbf{0} \right\}$$

with the inequality assumed coordinatewise. This is the case faced by individual investors, but not necessarily by financial institutions, which again may be subject to alternative constraints. In their case, short sales are usually allowed but too much exposure on particular assets is avoided. Hence, a more general objective can be used:

$$\mathbf{w}_{QP} = \arg\max_{\mathbf{w}} \left\{ \mathbf{w}^T \boldsymbol{\mu} - \frac{\kappa}{2} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \mid \mathbf{1}^T \mathbf{w} = 1,\ \mathbf{w}^l \leq \mathbf{w} \leq \mathbf{w}^u \right\}$$

where once again, the inequalities are assumed coordinatewise. Finally, further constraints can be imposed on linear combinations of the portfolio

[3]Optimization problems involving uncertainty in either the input parameters or the constraints are often approached through stochastic programming techniques (see, for example, Birge and Louveaux (1997) or Kall and Wallace (1999)). The shrinkage methods mentioned here together with the research undertaken in this thesis offer alternative solutions.

weights resulting in

$$\mathbf{w}_{QP} = \arg\max_{\mathbf{w}} \left\{ \mathbf{w}^T\boldsymbol{\mu} - \frac{\kappa}{2}\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \mid \mathbf{R}\mathbf{w} = \mathbf{r},\ \mathbf{w}^l \le \mathbf{w} \le \mathbf{w}^u \right\}. \tag{2.2}$$

Here, $\mathbf{R}$ can be thought of as a restriction matrix, and thus the system of equations $\mathbf{R}\mathbf{w} = \mathbf{r}$ will also include the full investment constraint $\mathbf{1}^T\mathbf{w} = 1$. It is obvious that very general optimization scenarios can be included, such as investing a specified percentage of a portfolio in a particular industry.[4]

Jagannathan and Ma (2002) raise an important issue concerning constrained optimization. They show that imposing non-negativity constraints on the portfolio weights and using the sample covariance matrix $\mathbf{S}$ as an estimator for $\boldsymbol{\Sigma}$ is equivalent to reducing the large elements in $\mathbf{S}$ and performing the optimization without any restrictions. This is effectively equivalent to shrinkage of the covariance matrix towards specified targets, which we focus upon in the next section.

2.3.2 Shrinkage Estimators

One way of achieving stability in the estimated portfolio is through shrinkage. The sample moments are prone to estimation error, hence a better alternative is to shrink them towards pre-determined targets. These targets may reflect some prior beliefs or knowledge or simply, when combined, may produce desirable portfolios. Furthermore, Victoria-Feser (2000) argues that robust estimators for both the location vector and the covariance matrix are also necessary in the case of deviations from the Normality assumption. We survey some of the proposed shrinkage methods here.

James-Stein Estimator

We start with a rather unusual but very interesting result. Stein (1956) proved that if the $N \times 1$ vector $\mathbf{x} \sim \mathcal{MN}(\boldsymbol{\mu}, \mathbf{I})$ then, for $N \ge 3$, the usual

[4]Moving away from linear restrictions, constraints can also be imposed on the number of assets on which investment should be made although such considerations are beyond the scope of this thesis.

multivariate sample mean $\bar{\mathbf{x}}$ is not admissible under the quadratic error loss function. Efron and Morris (1976) generalised the inadmissibility result for the case $\mathbf{x} \sim \mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a known covariance $\boldsymbol{\Sigma}$ under the loss function

$$L(\bar{\mathbf{x}}, \boldsymbol{\mu}) = (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}).$$

As a result of James and Stein (1961), Jorion (1986) argues that the estimator:

$$\hat{\boldsymbol{\mu}}_{JS} = (1 - c_{JS})\,\bar{\mathbf{x}} + c_{JS}\,\hat{\mu}\,\mathbf{1}$$

with

$$c_{JS} = \min\left\{1, \frac{(N-2)/T}{(\bar{\mathbf{x}} - \hat{\mu}\,\mathbf{1})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \hat{\mu}\,\mathbf{1})}\right\}$$

has uniformly lower risk compared to $\bar{\mathbf{x}}$. It is interesting that $\hat{\mu}$ may be chosen to be any point and still the estimator will be better than $\bar{\mathbf{x}}$.

Based on this, the *James-Stein Estimator* $\hat{\mathbf{w}}_{JS}$ is given by

$$\hat{\mathbf{w}}_{JS} = \frac{\mathbf{S}^{-1} \hat{\boldsymbol{\mu}}_{JS}}{\mathbf{1}^T \mathbf{S}^{-1} \hat{\boldsymbol{\mu}}_{JS}}$$

where $\hat{\boldsymbol{\mu}}_{JS}$ is defined as above with $\boldsymbol{\Sigma}$ estimated by $\mathbf{S}$ and

$$\hat{\mu} = \frac{1}{N} \sum_{j=1}^{N} \bar{x}_j\,.$$

The argument behind this estimator is that the target, $\hat{\mu}\,\mathbf{1}$, towards which shrinkage is applied will be less susceptible to estimation error than $\bar{\mathbf{x}}$. At the same time, to assess the need for shrinkage, the variability in the data is used: if the diagonal elements of $\mathbf{S}$ are small, then the second term in the shrinkage factor c_{JS} will be small and as a result very little shrinkage will be applied. On the other hand, if the data exhibit a large degree of variability, the shrinkage estimator of the mean may simply be $\hat{\mu}\,\mathbf{1}$ (i.e. $c_{JS} = 1$) which will protect the investor from unstable estimates.

It is also important to note that as we shrink the location vector towards $\hat{\mu}\,\mathbf{1}$, we effectively shrink non-linearly the estimated portfolio $\hat{\mathbf{w}}_{JS}$ towards the sample global minimum variance portfolio:

$$\hat{\mathbf{w}}_{GMV} = \frac{\mathbf{S}^{-1}\mathbf{1}}{\mathbf{1}^T \mathbf{S}^{-1}\mathbf{1}}.$$

This is compatible with our findings in Figure 2.2 and 2.3 illustrating how portfolios towards the bottom of the efficient frontier (i.e. close to the global minimum variance portfolio) exhibit lower out-of-sample variability compared to ones closer to the tangency portfolio.

Ledoit Estimator

The James-Stein methodology assumes that the covariance matrix is adequately estimated by the data. This may not be the case under all circumstances, especially when faced with large ($N \geq T$) covariance matrices. Ledoit (1995) introduces a covariance estimator which, on the one hand, is well-conditioned and therefore does not augment estimation error when inverted, and, on the other hand, performs asymptotically better than the sample covariance matrix $\mathbf{S}$ when the number of assets converges to infinity as well. This estimator is a convex linear combination of the sample covariance matrix and the identity matrix.

Rather than using a multiple of the identity matrix as a shrinkage target, Ledoit (1997) proposes the use of the single-index covariance matrix. As a result the *Ledoit Estimator* for the covariance matrix is given by:

$$\hat{\mathbf{\Sigma}}_{Ld} = \frac{\alpha}{T}\mathbf{F} + \left(1 - \frac{\alpha}{T}\right)\mathbf{S}$$

where, α is the shrinkage intensity, T is the sample size and $\mathbf{F}$ is the covariance matrix obtained by the single factor model (Sharpe, 1963). This leads to the estimated portfolio

$$\hat{\mathbf{w}}_{Ld} = \frac{\hat{\mathbf{\Sigma}}_{Ld}^{-1}\bar{\mathbf{x}}}{\mathbf{1}^T\hat{\mathbf{\Sigma}}_{Ld}^{-1}\bar{\mathbf{x}}}$$

The shrinkage factor α is chosen to be asymptotically optimal at minimising a quadratic measure of distance between the estimated and true covariance matrices. More specifically, α is chosen to minimize the asymptotic expectation of the loss function $||\hat{\mathbf{\Sigma}}_{Ld} - \mathbf{\Sigma}||^2$ where $||Z||^2$ denotes the squared Frobenius norm of Z.

It is interesting to note that using the Ledoit Estimator has similarities with the standard k-factor model in the finance literature: at the one end of the spectrum lies the single-factor covariance matrix $\mathbf{F}$ whereas at the other, the estimator becomes the sample covariance matrix, $\mathbf{S}$, which is equivalent to the N-factor model. By using this shrinkage estimator, a k-factor model is assumed to be correct with $1 \leq k \leq N$. However, using $\hat{\mathbf{\Sigma}}_{Ld}$ has the added advantage of not having to pre-determine the number of factors, k, which would be necessary otherwise.

Britten-Jones Estimator

The Ledoit approach is extended by Britten-Jones (2000). First of all, with respect to the covariance matrix, an additional target matrix is used. The new *Britten-Jones Estimator* for the covariance matrix becomes:

$$\hat{\mathbf{\Sigma}}_{BJ} = \alpha_1 \mathbf{S} + \alpha_2 \mathbf{I} + \alpha_3 \mathbf{J}$$

where $\mathbf{J}$ is a matrix of ones. The expected sum of squared deviations between the elements of $\hat{\mathbf{\Sigma}}_{BJ}$ and the true covariance matrix $\mathbf{\Sigma}$ is minimised to estimate the coefficients $(\alpha_1, \alpha_2, \alpha_3)$. A similar procedure is employed for the shrinkage estimator of the sample mean:

$$\hat{\boldsymbol{\mu}}_{BJ} = \beta_1 \bar{\mathbf{x}} + \beta_2 \mathbf{1}$$

with the coefficients (β_1, β_2) once again estimated by minimising the expected sum of squared errors. Plugging-in both $\hat{\mathbf{\Sigma}}_{BJ}$ and $\hat{\boldsymbol{\mu}}_{BJ}$ in equation (1.2) results in the estimated tangency portfolio while, $\hat{\mathbf{\Sigma}}_{BJ}$ and equation (1.3) are used for the minimum variance portfolio.

Frost-Savarino Estimator

A unifying approach to shrinkage is provided through the Bayesian framework and the posterior moments of the joint predictive distribution of returns. Under the assumption of multivariate Normality of returns, a Normal-Wishart conjugate prior leads to a multivariate t-distributed predictive density for the asset returns. Frost and Savarino (1986) specify an informative prior based on identical means, variances and pairwise correlation coefficients across all asset returns. Consistency (or lack of it) of the observed sample with the informative prior distribution determines the optimal shrinkage intensity.

More specifically, using prior parameters $\mathbf{\Omega}^{-1}, \nu, \boldsymbol{\mu}_0$ and τ, the Normal-Wishart conjugate prior is denoted by:

$$f_{NW}\left(\boldsymbol{\mu}, \mathbf{\Lambda} \,|\, \mathbf{\Omega}^{-1}, \nu, \boldsymbol{\mu}_0, \tau\right) = f_N\left(\boldsymbol{\mu} \,|\, \boldsymbol{\mu}_0, (\tau\mathbf{\Lambda})^{-1}\right) \times f_W\left(\mathbf{\Lambda} \,|\, (\nu\mathbf{\Omega})^{-1}, \nu\right)$$

where $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$. The strength of belief in the prior values $\mathbf{\Omega}^{-1}$ and $\boldsymbol{\mu}_0$ is adjusted by the prior parameters ν and τ respectively. Based on this assumption and a multivariate normal distribution of returns, the predictive density of returns is a multivariate t-distribution involving the posterior estimates of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, denoted by $\boldsymbol{\mu}_{FS}$ and $\mathbf{\Sigma}_{FS}$. These depend on belief weights ω_τ and ω_ν: as ω_τ increases, the posterior mean $\boldsymbol{\mu}_{FS}$ approaches the prior value $\boldsymbol{\mu}_0$ and similarly for $\mathbf{\Sigma}$.

Rather than using a full Bayesian approach, Frost and Savarino (1986) use point estimates for the prior parameters $\mathbf{\Omega}^{-1}, \nu, \boldsymbol{\mu}_0$ and τ. Under the assumption of equal means, variances and pairwise correlation coefficients across all asset returns, the values $\mathbf{\Omega}^{-1}$ and $\boldsymbol{\mu}_0$ are equated to their maximum likelihood sample estimates $\hat{\mathbf{\Omega}}^{-1}$ and $\hat{\boldsymbol{\mu}}_0 \propto \mathbf{1}$. This means that the shrinkage target for each element in the posterior mean is the overall mean in a similar fashion with the James-Stein estimator whereas the shrinkage target for the covariance matrix will be a linear combination of two matrices: a matrix proportional to the identity matrix and a matrix proportional to a matrix of ones, similarly to the Britten-Jones procedure. Empirical Bayes estimates of the remaining two prior parameters (ν and τ) are used by maximizing the likelihood function conditional on $\hat{\mathbf{\Omega}}^{-1}$ and $\hat{\boldsymbol{\mu}}_0$.

By plugging-in the chosen prior parameters, Frost and Savarino (1986) obtain the posterior estimates $\hat{\boldsymbol{\mu}}_{FS}$ and $\hat{\mathbf{\Sigma}}_{FS}$ which can be used to find optimal tangency and minimum variance portfolios as before.

2.3.3 Bayesian Regression

An alternative way of viewing the portfolio optimization problem was introduced by Jobson and Korkie (1983). Consider an artificial regression of the assets returns on a vector of ones:

$$\mathbf{1} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{2.3}$$

where $\mathbf{X}$ is a $T \times N$ matrix of excess asset returns. The least squares estimator for $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{1}.$$

It is easy to show (Britten-Jones, 1999) that

$$\frac{\hat{\boldsymbol{\beta}}}{\mathbf{1}^T\hat{\boldsymbol{\beta}}} = \frac{\mathbf{S}^{-1}\bar{\mathbf{x}}}{\mathbf{1}^T\mathbf{S}^{-1}\bar{\mathbf{x}}} = \hat{\mathbf{w}}_{SM}$$

and hence the sample moments estimator can be recovered from this projection. Britten-Jones (2000) develops Bayesian inference procedures for the projection coefficients $\boldsymbol{\beta}$ based on the multivariate normal likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the prior distribution $\pi(\boldsymbol{\beta}) \sim \mathcal{MN}(\boldsymbol{\beta}_0, \mathbf{C}_0)$. The likelihood can be written as a function of $\boldsymbol{\beta}$ where the projection coefficients $\boldsymbol{\beta}$ minimise $E\left[\mathbf{u}^T\mathbf{u}\right] = E\left[(\mathbf{1} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{1} - \mathbf{X}\boldsymbol{\beta})\right]$ and of the nuisance parameter $\mathbf{M} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$. The likelihood is then factorised into one term involving two parameters ($\boldsymbol{\beta}$ and $\kappa^2 = 1 - \boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}$) and another term involving just $\mathbf{M}$. Britten-Jones (2000) follows Cox (1975) in omitting the nuisance parameter term and uses the partial likelihood $L^P(\boldsymbol{\beta}, \kappa^2)$ (rather than the full likelihood) to obtain the posterior distribution $\boldsymbol{\beta} \,|\, \mathbf{X} \sim \mathcal{MN}(\mathbf{m}, \mathbf{V})$ where:

$$\mathbf{m} = \left(\mathbf{C}_0^{-1} + \kappa^{-2}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{C}_0^{-1}\boldsymbol{\beta}_0 + \kappa^{-2}\mathbf{X}^T\mathbf{X}\,\hat{\boldsymbol{\beta}}\right)$$

and

$$\mathbf{V} = \left(\mathbf{C}_0^{-1} + \kappa^{-2}\mathbf{X}^T\mathbf{X}\right)^{-1},$$

with κ^2 assumed known. Further investigation shows that the cost of using the partial rather than the full likelihood function is not very important. The excluded term involves the *length* of the vector $\boldsymbol{\beta}$ and not its *direction* which, when the portfolio weights $\mathbf{w}$ are recovered, will be preserved.

Finally, similarly to the Ledoit Estimator, the Bayesian Regression estimator also works when assets outnumber the observations and the sample covariance matrix is singular. This is because the second factor needed for the evaluation of the posterior mean $\mathbf{m}$ now becomes $\left(\mathbf{C}_0^{-1}\boldsymbol{\beta}_0 + \kappa^{-2}\mathbf{X}^T\mathbf{1}\right)$ and does not involve the inversion of $\mathbf{X}^T\mathbf{X}$.

The one significant drawback of this method (and, in fact, of most of the shrinkage estimators) would be encountered at the normalisation stage. Once the optimal $\boldsymbol{\beta}$ have been chosen, we recover the portfolio weights by imposing the full investment constraint, i.e. by dividing each of the elements in $\boldsymbol{\beta}$ by $\mathbf{1}^T\boldsymbol{\beta}$. This may still result in extreme portfolio weights when the sum of the elements in $\boldsymbol{\beta}$ is close to 0. We tackle this problem and propose some solutions in the next chapter.

Chapter 3

Bootstrap Portfolio Selection

The previous chapter highlighted the significant uncertainty associated with points on the estimated efficient frontier and, more specifically, the sample tangency portfolio. It is therefore of great importance to account for this uncertainty when choosing an estimator for the tangency portfolio. In this chapter, we propose overcoming this problem through the use of robust measures of location and resampling techniques.

3.1 Resampling Methods

Ever since computational power has been easily available to statisticians, the use of resampling methods increased considerably. Resampling effectively generates additional datasets from the original dataset with the objective of extracting more information about the properties of a statistic or quantity of interest. Named after the legendary story of Baron Munchausen, who thought of pulling himself out of the bottom of a lake by his bootstraps, bootstrapping is the main resampling method. In this section we briefly introduce the *Frequentist* and the *Bayesian bootstrap.*

3.1.1 The Frequentist Bootstrap

The notion of the bootstrap as a technique of drawing inferences about unknown parameters was introduced by Efron (1979). Suppose we have T realisations of a random variable $\mathbf{x}$ in a sample $\mathbf{x}_1, \ldots, \mathbf{x}_T$ and we are interested in a statistic $\mathbf{t}(\mathbf{x}_1, \ldots, \mathbf{x}_T)$. We can take a simple random sample with replacement of size T from the original sample, and calculate the value of the statistic $\mathbf{t}_i(\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_T^{(i)})$, where $\left\{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_T^{(i)}\right\}$ denotes the ith resample. Taking all possible bootstrap resamples and evaluating the statistic $\mathbf{t}_i$ each time, we obtain the bootstrap distribution of this statistic (see, for example, Efron and Tibshirani (1993)).

In an equivalent manner, bootstrap inference can be made from paired observations rather than a single sample. In general, let us assume the regression framework, whereby the vector of explanatory variables $\mathbf{x}_t$ is associated with the response variable y_t, for each $t = 1, \ldots, T$ through, for example,

$$y_t = \boldsymbol{\beta}^T \mathbf{x}_t + \epsilon_t.$$

Rather than the ordinary least squares estimator (or any other estimator $\hat{\boldsymbol{\beta}}$), the frequentist bootstrap methodology proposes taking a simple random sample with replacement of size T from $(y_t, \mathbf{x}_t)$ pairs. As before, the sampling is repeated until B bootstrap resamples are obtained. For each resample $\left\{\left(y_1^{(i)}, \mathbf{x}_1^{(i)}\right), \ldots, \left(y_T^{(i)}, \mathbf{x}_T^{(i)}\right)\right\}$ a new estimator, $\hat{\boldsymbol{\beta}}^{(i)}$, is calculated. As $B \to \infty$ the distribution of the set of estimators $\left\{\hat{\boldsymbol{\beta}}^{(i)}\right\}$ approaches the bootstrap distribution of $\hat{\boldsymbol{\beta}}$.

3.1.2 The Bayesian Bootstrap

Before proceeding to explain the robust estimator of portfolio weights, we examine here an alternative way of generating additional resamples in order to account for between-sample variability, namely the Bayesian bootstrap (Rubin, 1981). Rather than taking resamples from the observed density function and hence assigning probabilities of $\frac{1}{T}$ for each sample point, Rubin

(1981) proposes using the posterior probability for each of the observed data points, centred around $\frac{1}{T}$ but at the same time exhibiting variability.

More specifically, a Bayesian Bootstrap resample is obtained in the following way: $(T-1)$ Uniform (0, 1) random numbers $u_1, u_2, \ldots, u_{T-1}$ are generated, arranged in ascending order and used to evaluate the gaps $g_i = u_{(i)} - u_{(i-1)}$ for $i = 1, \ldots, T$ with $u_{(0)} = 0$ and $u_{(T)} = 1$. Each of the elements of the $T \times 1$ vector $\mathbf{g} = (g_1, g_2, \ldots, g_T)^T$ forms the probabilities attached with the respective observed data vector $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The Bayesian Bootstrap resample is generated by sampling from the original data sample and selecting data point $\mathbf{x}_t$ with probability g_t for $t = 1, \ldots, T$ until a sample of size T is obtained.

3.1.3 The Bootstrap Distribution of an Estimator

Suppose now that we are interested in optimising a (possibly non-linear) function of the unknown theoretical values $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Examples of such loss functions include the Sharpe Ratio, the expected utility or the Value-at-Risk of a portfolio. Since quantities such as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, the optimization objective can be approximated as a function of their sample counterparts, $\bar{\mathbf{x}}$ and $\mathbf{S}$, and therefore, as a function of the data $f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$. These functions are defined over the space of portfolio weights and have usually a single optimum. The conventional sample estimator is the vector of portfolio weights which optimises $f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ i.e.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \{f(\mathbf{x}_1, \ldots, \mathbf{x}_T)\}$$

subject to possible constraints such as $\mathbf{1}^T\mathbf{w} = 1$, $\mathbf{w} \geq \mathbf{0}$ etc. In the case where a model is used (such as the one in equation (2.3)) the estimator can be written as:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \{g\left[(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)\right]\}$$

subject to similar constraints.

If the sampling distribution of the estimator is unknown or difficult to derive, one may consider the *bootstrap distribution of the estimator*, i.e. the set of values $\left\{\hat{\mathbf{w}}^{(i)} : i = 1, \ldots, B\right\}$ where

$$\hat{\mathbf{w}}^{(i)} = \arg\min_{\mathbf{w}} \left\{ f\left(\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_T^{(i)}\right)\right\}$$

and $\left(\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_T^{(i)}\right)$ denotes the ith bootstrap resample. The bootstrap distribution of $\hat{\mathbf{w}}$ is usually considered as a means of assessing the estimator's uncertainty in terms of its standard error. Moreover, using bootstrap replications one is able to obtain estimates of an estimator's bias. We will use in this chapter an alternative viewpoint: using the bootstrap distribution $\left\{\hat{\mathbf{w}}^{(i)} : i = 1, \ldots, B\right\}$ to improve, on average, the *accuracy* of an estimator.

This procedure has its roots in what Breiman (1996) calls "bootstrap aggregating" or *bagging.* Breiman (1996) proposes the use of the mean of the bootstrap distribution as a better estimator compared to the original sample estimator. The procedure is especially useful when the original estimator is unstable, in the sense that small changes in the sample may lead to large changes in the estimator.

3.2 Robust Estimation of Location

In many cases the bootstrap distribution of a statistic, such as the sample tangency portfolio, exhibits heavy tails. On the one hand, this observation points towards the instability of the statistic under investigation. In other words, the statistic would exhibit large variability had other datasets been obtained even if the unknown data-generating process remained the same. On the other hand, this prompts us to select more robust estimation methods. Bootstrap resampling allows us to approximate the sampling distribution of the statistic by providing alternative samples that could have been encountered.

3.2.1 The Multivariate L_1 Median

One way of "averaging" over all these samples is to use a robust measure of location, since a mean-based point estimate is likely to severely fluctuate between samples.[1] A frequently used, robust estimator of location, is a multivariate analogue of the median. Since a multidimensional cloud of data does not possess a natural ordering,[2] in the sense that univariate data can be ranked, the univariate median has many multivariate analogues, most of them surveyed by Small (1990) and Chaudhuri (1992).

We will be using the multivariate L_1 median, first considered by Weber (1909) as the solution to a problem originating in Location Theory. A company has B customers located at co-ordinates $\hat{\mathbf{w}}^{(1)}, \ldots, \hat{\mathbf{w}}^{(B)}$ on an N-dimensional space. It wishes to choose an appropriate location (without any co-ordinate constraints) for its warehouse to service its B clients. Weber (1909) proposed the point minimising the total transportation costs and assuming that the costs are proportional to the Euclidean distance between a customer and the warehouse, he defined the *multivariate L_1 median* (also known as *mediancentre* (Gower, 1974) or *spatial median* (Brown, 1983)) of $\hat{\mathbf{w}}^{(1)}, \ldots, \hat{\mathbf{w}}^{(B)}$ as a point $\tilde{\mathbf{w}} \in \mathbb{R}^N$ such that

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{B} \|\hat{\mathbf{w}}^{(i)} - \mathbf{w}\|.$$

Note that for $N = 1$, $\tilde{\mathbf{w}}$ reduces to the univariate median while Milasevic and Ducharme (1987) show that for $N \geq 2$, the multivariate L_1 median is unique.

The next step in evaluating the usefulness of a robust estimator of location such as the L_1 median, concerns the estimator's characteristics. One such important characteristic is the *breakdown point*, which is roughly defined as the maximum proportion of contamination introduced in the data before

[1]This is the main difference with "bagging" (Breiman, 1996) which uses the mean of the bootstrap distribution.

[2]One way of ranking and viewing multivariate data is associated with the notion of *data depth* (see, for example, Liu et al. (1999)).

the estimator can be forced to take arbitrary values. For example, in a univariate sample of size B, the mean has a breakdown point of $\frac{1}{B}$ since only one observation need be altered to force the sample mean to take an arbitrary value. The univariate (L_1) median has a much higher breakdown of $\frac{1}{2}$, since 50% of the sample points need to be forced on one side of the median before it takes arbitrary values. In this sense, the multivariate analogue L_1 median has a similarly high breakdown point of $\frac{1}{2}$ (Lopuhaä and Rousseeuw, 1991) thus rendering the estimator very robust especially when estimating the location vector of a heavy-tailed multivariate distribution.

A further property to consider is that of *affine equivariance*. In the same sense that the univariate median is equivariant under monotonic transformations of the sample points, an analogously similar property is desirable in the multivariate setting. Small (1990) explains how the class of monotonic transformations on the line can be extended in higher dimensions with the equivariance property of the L_1 median preserved. However, this is not the case for all transformations. In fact, the L_1 median is generally not affine equivariant and for this reason, it has not been as widely used as one might have expected. Nevertheless, it is equivariant under translations and orthogonal transformations or rotations which in some cases is adequate.

3.2.2 An Affine Equivariant Median

Before proceeding, we pause here to consider an alternative estimator of multivariate location proposed by Hettmansperger and Randles (2002). They introduce an algorithm of finding an estimator for the location vector based on the transformation-retransformation technique of Chakraborty et al. (1998). The proposed multivariate median has the property of affine equivariance and combines the aforementioned L_1 median with an M-estimator of scatter proposed by Tyler (1987).

The affine equivariant median estimator of location computed from a set of points $\hat{\mathbf{w}}^{(1)}, \ldots, \hat{\mathbf{w}}^{(B)}$ is given by:

$$\tilde{\mathbf{w}}_{AE} = \hat{\mathbf{A}}^{-1} \times \arg\min_{\mathbf{w}} \sum_{i=1}^{B} \hat{\mathbf{A}} \left\|\hat{\mathbf{w}}^{(i)} - \mathbf{w}\right\|$$

where $\left(\hat{\mathbf{A}}^T\hat{\mathbf{A}}\right)^{-1}$ is the M-estimator of scatter. Based on the equivariance of the L_1 median under orthogonal transformations, it is obvious that when $\hat{\mathbf{A}}$ is an orthogonal transformation, the affine equivariant median $\tilde{\mathbf{w}}_{AE}$ will coincide with the aforementioned L_1 median. In such a case $\left(\hat{\mathbf{A}}^T\hat{\mathbf{A}}\right)^{-1}$ would simply be the identity matrix. Working backwards, we can infer that if the variance in each dimension of $\hat{\mathbf{w}}^{(1)}, \ldots, \hat{\mathbf{w}}^{(B)}$ is similar and the correlations between the elements are small then the results of either of the two estimators will be similar as well, a fact that was confirmed by simulations.

In conclusion, we are interested in a robust estimator of location of the approximate bootstrap distribution of $\hat{\mathbf{w}}$. In this setting, on the one hand all elements of $\hat{\mathbf{w}}$ will be measured on the same scale without the need of different re-scaling in different dimensions. Hence, it is likely that their respective variances will be similar. On the other hand, by construction, correlations between elements of $\hat{\mathbf{w}}$ will not be zero since $\mathbf{1}^T\hat{\mathbf{w}} = 1$ unless there is a sufficiently large number of assets available. Nevertheless, experimental work suggests that using the multivariate L_1 median, $\tilde{\mathbf{w}}$, will be adequate for our purposes.

3.3 The Bootstrap L_1 Median Portfolio

Having proposed using the multivariate L_1 median of an estimator's bootstrap distribution as an improvement on the estimator's performance, we formally present it here:

Definition 1 *Let $\hat{\mathbf{w}}^{(i)}$ denote an estimator $f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ applied to the ith bootstrap resample $\left(\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_T^{(i)}\right)$ for $i = 1, \ldots, B$. Then, the bootstrap L_1 median portfolio estimator is defined as:*

$$\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{B} \sum_{i=1}^{B} \sqrt{(\hat{\mathbf{w}}^{(i)} - \mathbf{w})^T(\hat{\mathbf{w}}^{(i)} - \mathbf{w})} \right\} \tag{3.1}$$

Possible choices for the function $f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of the data include the numerous estimators that were introduced in the previous chapter. Furthermore if, for each i, the vectors $\hat{\mathbf{w}}^{(i)}$ satisfy both equality and inequality constraints, then these properties will also describe the vector $\tilde{\mathbf{w}}$ as the following theorem states:

Theorem 1 *Define $\tilde{\mathbf{w}}$ as in equation (3.1) and let vectors $\hat{\mathbf{w}}^{(1)}, \hat{\mathbf{w}}^{(2)}, \ldots, \hat{\mathbf{w}}^{(B)}$ satisfy linear constraints $\mathbf{R}\hat{\mathbf{w}}^{(i)} = \mathbf{r}$ and co-ordinatewise inequalities $\mathbf{w}^l \leq \hat{\mathbf{w}}^{(i)} \leq \mathbf{w}^u$ for $i = 1, \ldots, B$. Then (a) $\mathbf{R}\tilde{\mathbf{w}} = \mathbf{r}$ and (b) $\mathbf{w}^l \leq \tilde{\mathbf{w}} \leq \mathbf{w}^u$.*

To prove Theorem 1(a), we consider Figure 3.1. The diagram illustrates how points $A, B, \ldots, E$ lie on the hyperplane created by the linear constraints $\mathbf{R}\hat{\mathbf{w}} = \mathbf{r}$. For illustrative purposes, the $N \times 1$ vectors $\hat{\mathbf{w}}^{(1)}, \hat{\mathbf{w}}^{(2)}, \ldots, \hat{\mathbf{w}}^{(B)}$ are represented by $A, B, \ldots, E$. Suppose that the proposed multivariate L_1

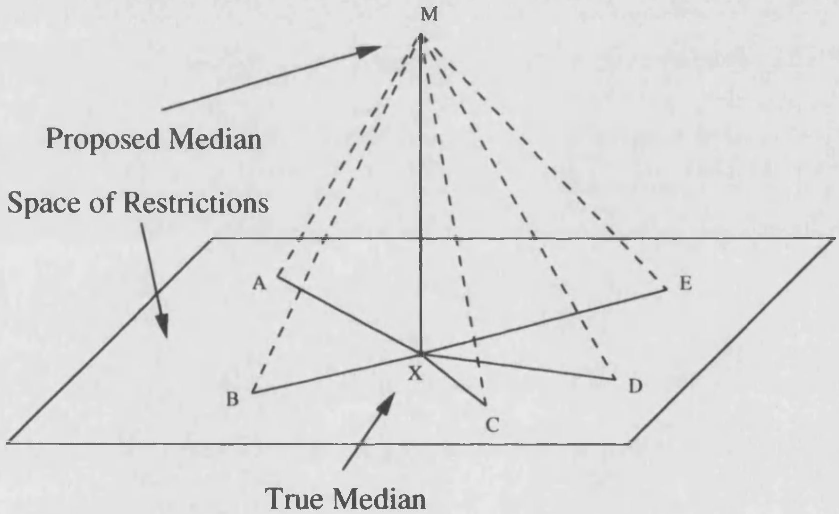

Figure 3.1: Multivariate L_1 Median and Equality Constraints

median $\tilde{\mathbf{w}}$ (Point M) did not belong on the hyperplane of restrictions i.e. that $\mathbf{R}\tilde{\mathbf{w}} \neq \mathbf{r}$ and let $\tilde{\mathbf{w}}_x$ be the projection of $\tilde{\mathbf{w}}$ on $\mathbf{R}\hat{\mathbf{w}} = \mathbf{r}$. Then by the Pythagoras' Theorem:

$$\underbrace{\sum_{i=1}^{B} \sqrt{(\hat{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}})^T(\hat{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}})}}_{|AM|+\ldots+|EM|} =$$

$$= \underbrace{\sum_{i=1}^{B} \sqrt{(\hat{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}}_x)^T(\hat{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}}_x) + (\tilde{\mathbf{w}} - \tilde{\mathbf{w}}_x)^T(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}_x)}}_{\sqrt{|AX|^2+|MX|^2}+\ldots+\sqrt{|EX|^2+|MX|^2}} \geq$$

$$\geq \underbrace{\sum_{i=1}^{B} \sqrt{(\hat{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}}_x)^T(\hat{\mathbf{w}}^{(i)} - \tilde{\mathbf{w}}_x)}}_{|AX|+\ldots+|EX|}$$

As a result,

$$\tilde{\mathbf{w}} \neq \arg\min_{\mathbf{w}} \left\{ \frac{1}{B} \sum_{i=1}^{B} \sqrt{(\hat{\mathbf{w}}^{(i)} - \mathbf{w})^T(\hat{\mathbf{w}}^{(i)} - \mathbf{w})} \right\}$$

unless $\tilde{\mathbf{w}} \equiv \tilde{\mathbf{w}}_x$ which means that $\tilde{\mathbf{w}}$ must satisfy $\mathbf{R}\tilde{\mathbf{w}} = \mathbf{r}$ as required by Theorem 1(a).

We now turn our attention to the proof of Theorem 1(b). We shall prove this for $N = 2$ for illustrative purposes but the proof can be extended to higher dimensions. Consider Figure 3.2. Suppose that points $A, B, \ldots, E$ form the vertices of the convex hull of $\hat{\mathbf{w}}^{(1)}, \hat{\mathbf{w}}^{(2)}, \ldots, \hat{\mathbf{w}}^{(B)}$. Let point M lie outside the convex hull and point X be the projection of M onto the nearest edge. Denote by $L(M)$ the sum of Euclidean distances from M to each of $A, B, \ldots, E$, i.e. $L(M) = |AM| + |BM| + \ldots + |EM|$. We need to show that $L(M) \geq L(X)$.

By the cosine rule:

$$\begin{aligned}
|AM| &= \sqrt{|AX|^2 + |MX|^2} \\
|BM| &= \sqrt{|BX|^2 + |MX|^2 - 2\,|BX||MX| \cos(\widehat{BXM})} \\
|CM| &= \sqrt{|CX|^2 + |MX|^2 - 2\,|CX||MX| \cos(\widehat{CXM})} \\
|DM| &= \sqrt{|DX|^2 + |MX|^2 - 2\,|DX||MX| \cos(\widehat{DXM})} \\
|EM| &= \sqrt{|EX|^2 + |MX|^2}
\end{aligned}$$

The angles $\widehat{BXM}, \widehat{CXM}$ and $\widehat{DXM}$ will be obtuse since by construction $\widehat{AXM}$ and $\widehat{EXM}$ are right angles. Hence, $\cos(\widehat{BXM}), \cos(\widehat{CXM})$ and $\cos(\widehat{DXM})$ will be negative and therefore

$$L(M) \geq \sqrt{|AX|^2 + |MX|^2} + \ldots + \sqrt{|EX|^2 + |MX|^2}$$

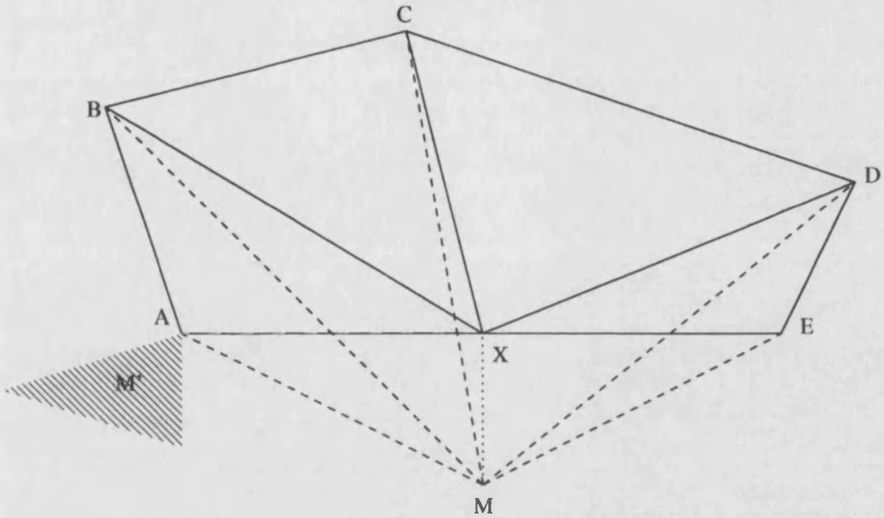

Figure 3.2: Multivariate L_1 Median and Convex Set

$$
\begin{aligned}
&\geq |AX| + \ldots + |EX| \\
&= L(X)
\end{aligned}
$$

In order to complete the proof, we need to consider the case of a point M' outside the convex hull but in an area where its nearest point in the convex hull is one of the vertices. Such is the case in the shaded area of Figure 3.2 for vertex A and equivalently for the remaining vertices. In such a case we need to show that $L(M') \geq L(A)$. We proceed as before with:

$$
\begin{aligned}
|BM'| &= \sqrt{|AB|^2 + |AM'|^2 - 2\,|AB||AM'|\cos(\widehat{BAM'})} \\
|CM'| &= \sqrt{|AC|^2 + |AM'|^2 - 2\,|AC||AM'|\cos(\widehat{CAM'})} \\
|DM'| &= \sqrt{|AD|^2 + |AM'|^2 - 2\,|AD||AM'|\cos(\widehat{DAM'})} \\
|EM'| &= \sqrt{|AE|^2 + |AM'|^2 - 2\,|AE||AM'|\cos(\widehat{EAM'})}
\end{aligned}
$$

For all possible points M' in the shaded area of Figure 3.2, the angles $\widehat{BAM'}$, $\widehat{CAM'}$, $\widehat{DAM'}$ and $\widehat{EAM'}$ will be obtuse or right angles (if M' lies on the boundary) hence $\cos(\widehat{BAM'})$, $\cos(\widehat{CAM'})$, $\cos(\widehat{DAM'})$, $\cos(\widehat{EAM'}) \leq 0$ and therefore

$$
L(M') \geq |AM'| + \sqrt{|AB|^2 + |AM'|^2} + \ldots + \sqrt{|AE|^2 + |AM'|^2}
$$

$$\begin{aligned} &\geq |AB| + \ldots + |AE| \\ &= L(A) \end{aligned}$$

which completes the proof.

Hence, for every point M outside the convex hull of $\hat{\mathbf{w}}^{(1)}, \hat{\mathbf{w}}^{(2)}, \ldots, \hat{\mathbf{w}}^{(B)}$ there is another point X for which the sum of Euclidean distances from the $\hat{\mathbf{w}}^{(i)}$'s is smaller. As a result, $\tilde{\mathbf{w}}$ will lie within the convex hull and thus satisfy the inequality constraints $\mathbf{w}^l \leq \tilde{\mathbf{w}} \leq \mathbf{w}^u$.

3.4 Discussion

Before proceeding to illustrate the behaviour of the L_1 median estimator via a simulated example, in this section we will be investigating some issues that arise from this procedure. Furthermore, we will be offering one explanation on the reasons behind its success while at the same time highlighting its limitations.

3.4.1 Further Points on $\tilde{\mathbf{w}}$

This procedure is similar in a way to the one adopted by Michaud (1998), where resampling techniques are used to improve portfolio estimation: an investor is faced with some historical data which are described by certain sample parameters. However, even if these sample parameters had been accurate estimates of the true population parameters, an alternative dataset, with different sample properties, could have been obtained. Therefore, our work is similar to Michaud (1998) in the sense that both attempt to account for those alternative datasets which could have been encountered. However, there are three significant differences: firstly Michaud (1998) uses Monte Carlo simulations from the original sample parameters (mean, variances and covariances) whereas our work is based on bootstrap resamples from the original dataset. Secondly, his work is based on sign-constrained portfolios whereas we allow for other equality or inequality constraints or even short sales. Finally, and most importantly, he uses the Monte Carlo simulations to assess the need for revising a portfolio, or even to find statistically equivalent portfolios based on simulated uncertainty regions. We use the bootstrap resamples to obtain a more robust estimate of a particular portfolio estimator.

One advantage of the bootstrap L_1 median estimator is that it does not make any strong distributional assumptions on the data-generating process. It is often the case that certain estimators are derived based on such assumptions. However, these are sometimes unrealistic or unproven and hence limit an estimator's applicability. On the contrary, $\tilde{\mathbf{w}}$ only assumes that the multivariate bootstrap distribution of the original estimator is unimodal. This will ensure that the L_1 median will be a reasonable estimator of location.

A further advantage of $\tilde{\mathbf{w}}$ over the conventional estimator is related to the shape and tails of the multivariate bootstrap distribution. We assume that the bootstrap distribution is an approximation of the between-sample behaviour of the estimator under consideration. Characteristics of the unknown sampling distribution will mostly be reflected in the bootstrap realisations. If the bootstrap resamples give rise to a distribution which is well-behaved then little will be gained by the use of the $\tilde{\mathbf{w}}$ estimator. However, in the more likely scenario whereby the distribution exhibits heavy tails (and therefore any deviation from symmetry is augmented), the L_1 median will provide a more robust measure of location, less affected by extreme, but nevertheless possible, bootstrap realisations of the original estimator.

One point that may attract adverse criticism concerns the computational aspect of the estimator. Creating bootstrap resamples, applying the same estimator to each resample before finally evaluating the multivariate L_1 median might sound computationally very demanding, especially in the case that no significant improvement is achieved by the $\tilde{\mathbf{w}}$ estimator. However, in empirical tests ($N = 10, T = 120$) a large number of resamples (between 100 and 500) was more than adequate for our objective. With the capabilities of modern computers this issue does not pose any significant problems.

3.4.2 Successes and Failures of $\tilde{\mathbf{w}}$

Empirical simulations indicate that the bootstrap L_1 median estimator performs well when compared to the original sample moments estimator. The reason and conditions under which this improvement is possible is the same as in the case of bootstrap aggregating.

Let us assume that the observations $\mathbf{x}_1, \ldots, \mathbf{x}_T$ form a realisation from an unknown distribution $\mathcal{F}$. Borrowing terminology from the classification literature where bagging is predominantly used, these observations constitute the training set $\mathcal{L}$. In order to simplify notation, we assume that we are interested in a univariate estimator $\hat{w}_{\mathcal{L}} = f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and define the aggregated bootstrap estimate as

$$\tilde{\mathbf{w}} = E_{\mathcal{L}}\left[\hat{w}_{\mathcal{L}}\right].$$

Intuitively, this means that $\tilde{w}$ would be the average $\hat{w}_{\mathcal{L}}$ over all possible training sets $\mathcal{L}$. Similarly, the average squared error of $\hat{w}_{\mathcal{L}}$ as an estimator of the true, unknown w is given by

$$e = E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - w)^2\right].$$

We can write

$$\begin{aligned} E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - w)^2\right] &= E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - \tilde{w} + \tilde{w} - w)^2\right] \\ &= E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - \tilde{w})^2 + (\tilde{w} - w)^2 + 2\left(\tilde{w} - w\right)\left(\hat{w}_{\mathcal{L}} - \tilde{w}\right)\right] \\ &= (\tilde{w} - w)^2 + E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - \tilde{w})^2\right]. \end{aligned}$$

This is the bias-variance formula which entails that the average (over all training sets) squared error of $\hat{w}_{\mathcal{L}}$ is decomposed into a squared bias term, $(\tilde{w} - w)^2$, and a variance term, $E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - \tilde{w})^2\right]$. Using the Cauchy-Schwarz inequality $E\left[Z^2\right] \geq E\left[Z\right]^2$ or equivalently that $E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - \tilde{w})^2\right] \geq 0$, we have

$$e \geq (\tilde{w} - w)^2$$

hence , if all possible training sets could have been observed, $\tilde{w}$ would be, on average, closer to the truth, w, than the original estimator $\hat{w}_{\mathcal{L}}$.

Of course, we only observe one training set, $\mathcal{L}$, but Breiman (1996) argues that we can mimic the true distribution $\mathcal{F}$ by creating bootstrap resamples $\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_T^{(i)}$ for $i = 1, \ldots, B$. For each resample the estimator $\hat{w}_{\mathcal{L}}^{(i)}$ is calculated and in practice the bootstrap aggregated estimate is the mean of the series $\left\{\hat{w}_{\mathcal{L}}^{(i)}\right\}$. In this chapter, we adopted a more robust measure of location and used

$$\tilde{w}_A = \arg\min_w \left\{ \frac{1}{B} \sum_{i=1}^{B} \|\hat{w}_{\mathcal{L}}^{(i)} - w\| \right\}.$$

According to Breiman (1996), if the distributional approximation is adequate, bootstrap aggregating will improve the performance of an unstable estimator $\hat{w}_{\mathcal{L}}$. (A more formal definition of instability can be found in Bühlmann and Yu (2001).)

The degree of improvement achieved depends on the stability of the original estimator and more specifically on the magnitude of $E_{\mathcal{L}}\left[(\hat{w}_{\mathcal{L}} - \tilde{w})^2\right]$, i.e. the variance of the estimator. If the estimator is accompanied by large uncertainty, as empirical evidence has indicated regarding the tangency portfolio, then bootstrap aggregating will help. On the other hand, if the original estimator is stable, as in the case of the global minimum variance portfolio, the bootstrap L_1 median portfolio will not necessarily improve performance.

Friedman and Hall (2000) explain the success of bagging by decomposing statistical estimators into linear and higher order terms. They argue that bootstrap aggregating preserves the linear part of an estimator but reduces its variability by replacing the higher order terms by empirical approximations to their expected values. In a sense, the higher degree of non-linearity an estimator exhibits, the more potential improvement bagging can result in.

3.5 A Simulated Example

In this section we illustrate how the multivariate bootstrap L_1 median (or its affine equivariant counterpart presented earlier) can improve two point estimators: the global minimum variance portfolio and the unconstrained sample tangency portfolio. We assume that simulated observations follow a

$\mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with:

$$\boldsymbol{\mu} = \frac{1}{12}(0.08, 0.06, 0.04, -0.01)^T$$

and

$$\boldsymbol{\Sigma} = \frac{1}{12}\,\mathrm{diag}\,(0.24^2, 0.18^2, 0.16^2, 0.15^2).$$

The optimal global minimum variance portfolio, $\mathbf{w}_{GMV}$, and the optimal tangency portfolio, $\mathbf{w}_T$, as given by equations (1.3) and (1.2) respectively for this example are:

$$\begin{aligned} \mathbf{w}_{GMV} &= (0.13, 0.23, 0.30, 0.34)^T \\ \mathbf{w}_T &= (0.32, 0.42, 0.36, -0.10)^T \end{aligned}$$

Based on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we simulate 1000 samples, each of size $T = 120$ (thus representing 10 years of "monthly" data points). For each dataset we estimate the sample estimates of $\mathbf{w}_{GMV}$ and $\mathbf{w}_T$. We also generate $B = 500$ bootstrap resamples from each simulated dataset and proceed to calculate both the bootstrap L_1 median and the affine equivariant median estimator for $\mathbf{w}_{GMV}$ and $\mathbf{w}_T$. An algorithm for each of the respective multivariate medians, based on Bedall and Zimmermann (1979) and on Hettmansperger and Randles (2002), was written in Ox 3.20 (Doornik, 2002) for our calculations.

We pause here to explain that the analysis presented below refers to the L_1 version since it emerged from our simulations that both medians gave rise to almost identical results. We elected to use the L_1 median because of computational speed. It is faster than the affine equivariant version since the algorithm for the latter uses the L_1 median of transformed data before applying the re-transformation and repeats the process until convergence. The corresponding figures and tables for the analysis of the affine equivariant median can be found in Appendix C.

The sample point estimators, denoted by $\hat{\mathbf{w}}$, and their L_1 bootstrap median counterparts, denoted by $\tilde{\mathbf{w}}$, were compared on the basis of their respective "true" Sharpe Ratios (using $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) i.e.

$$SR(\tilde{\hat{\mathbf{w}}}) = \frac{\tilde{\hat{\mathbf{w}}}^T \boldsymbol{\mu}}{\sqrt{\tilde{\hat{\mathbf{w}}}^T \boldsymbol{\Sigma} \tilde{\hat{\mathbf{w}}}}}$$

for any portfolio estimator $\tilde{\hat{\mathbf{w}}}$. The results of these analyses are presented below.

3.5.1 Global Minimum Variance Portfolio

As illustrated in Figure 2.3 of the previous chapter, the minimum variance portfolio does not exhibit much variability between samples. Furthermore, as seen in the upper panel of Figure 3.3 which illustrates the distribution of $SR(\hat{\mathbf{w}})$ between the simulated datasets, the resulting Sharpe Ratio distribution is fairly symmetric. The lower panel of Figure 3.3 shows the Sharpe

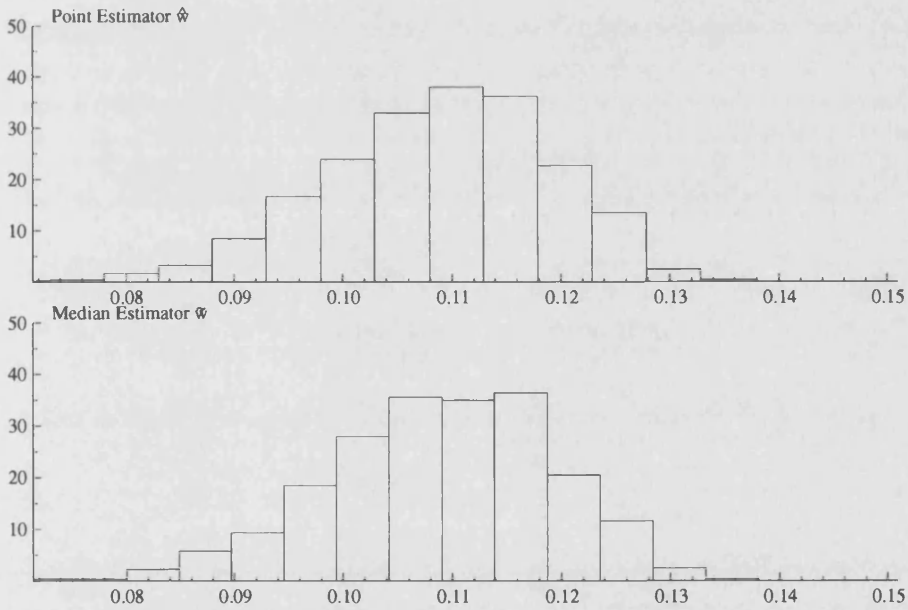

Figure 3.3: Distribution of Sharpe Ratio of GMV Portfolio

Ratio distribution with $\tilde{\mathbf{w}}$.[3] It is evident that the suggested portfolios under

[3]Figure C.1 in Appendix C shows the corresponding distribution with the affine equivariant median.

the L_1 median estimator give rise to Sharpe Ratios which are similar to the ones that would have been obtained under the usual sample point estimator, $\hat{\mathbf{w}}_{GMV}$.

Summary statistics for the two distributions are presented in Table 3.1 below.[4] These point towards similar conclusions regarding the performance of the two estimators. One further point of interest is the last row of the table which reveals how many times (out of 1000) each of the estimators gave rise to a higher "true" Sharpe Ratio. We can conclude that neither estimator dominates and that both have similar results.

	$SR(\hat{\mathbf{w}}_{GMV})$	$SR(\tilde{\mathbf{w}}_{GMV})$
Min	0.065	0.067
Q1	0.102	0.102
Mean	0.109	0.109
Median	0.109	0.109
Q2	0.116	0.116
Max	0.137	0.137
Res	484	516

Table 3.1: Summary Statistics for Estimators of $SR(\mathbf{w}_{GMV})$

A better graphical tool to compare the Sharpe Ratios achieved by each of the two estimators is given in Figure 3.4.[5] This plots the co-ordinates of the "true" Sharpe Ratios for each of the simulated datasets. If all points lay on the 45-degree line, the two estimators would have been exactly equivalent, as the pairs of Sharpe Ratios obtained for each dataset would have been the same. This is very nearly the case for the estimation of the minimum variance portfolio.

The above analysis exemplifies the argument that the bootstrap L_1 median estimator will perform at least as well as the conventional sample es-

[4]Table C.1 shows the corresponding statistics for the affine equivariant median.

[5]See also Figure C.2.

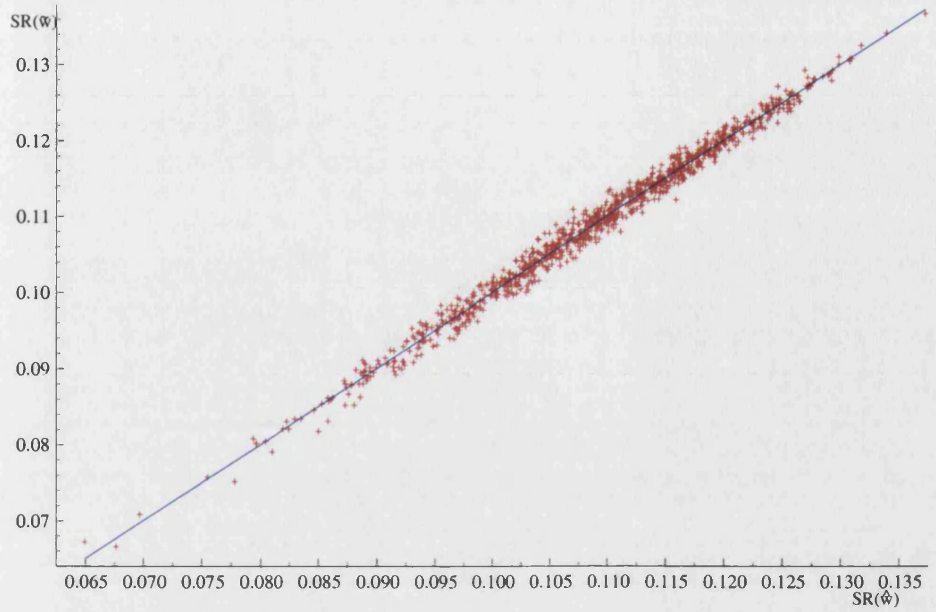

Figure 3.4: Sharpe Ratio of GMV Portfolio under Two Estimators

timator in cases where the estimator's distribution is well-behaved and the estimator itself is stable. This is the case for the global minimum variance portfolio. However, the true strengths of this procedure are illustrated in the following example.

3.5.2 Tangency Portfolio

The estimation of the sample tangency portfolio is notoriously difficult. One of the reasons for the problems encountered is the heavy-tailed distribution that the sample point estimator seems to follow. As a result, the "true" Sharpe Ratio obtained by using the sample moments estimator $\hat{\mathbf{w}}$, shown in the upper panel of Figure 3.5 is negatively skewed.[6] Although most values of $SR(\hat{\mathbf{w}})$ lie towards the maximum attainable value (which in this numerical example is given by $\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = 0.155$) there are nevertheless some datasets which give rise to estimators which in turn produce small (or even negative) Sharpe Ratios.

[6]See also Figure C.3.

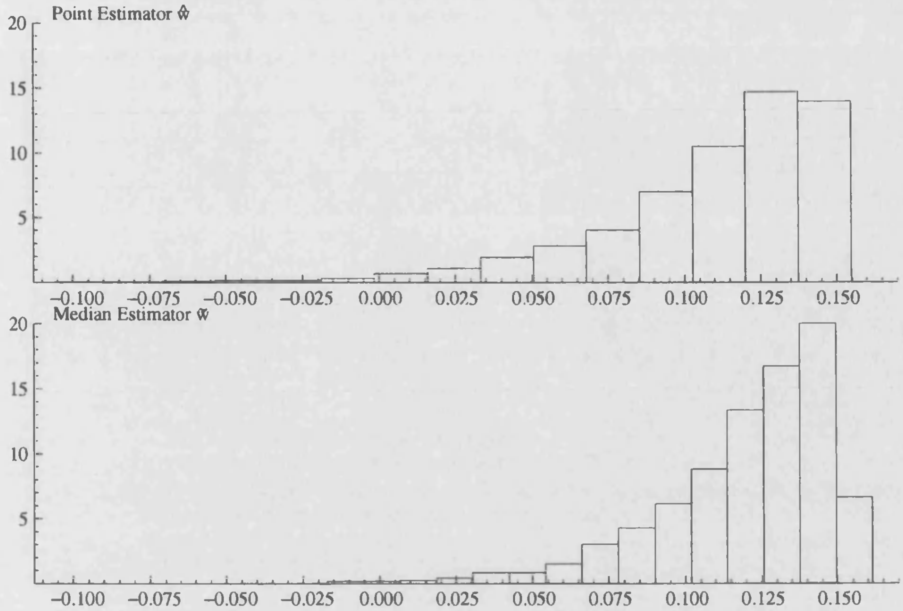

Figure 3.5: Distribution of Sharpe Ratio of Tangency Portfolio

The root of this problem can be traced back to the amount of between-sample variability exhibited by the sample tangency portfolio. Using the bootstrap L_1 median estimator goes some way to eradicate this problem by selecting a portfolio which is, on average, closer to the optimal portfolio. The distribution of the Sharpe Ratios obtained using this method are illustrated in the lower panel of Figure 3.5. Close inspection reveals how the L_1 median

	$SR(\hat{\mathbf{w}}_T)$	$SR(\tilde{\mathbf{w}}_T)$
Min	-0.102	-0.017
Q1	0.092	0.107
Mean	0.110	0.120
Median	0.119	0.127
Q2	0.137	0.142
Max	0.155	0.155
Res	271	729

Table 3.2: Summary Statistics for Estimators of $SR(\mathbf{w}_T)$

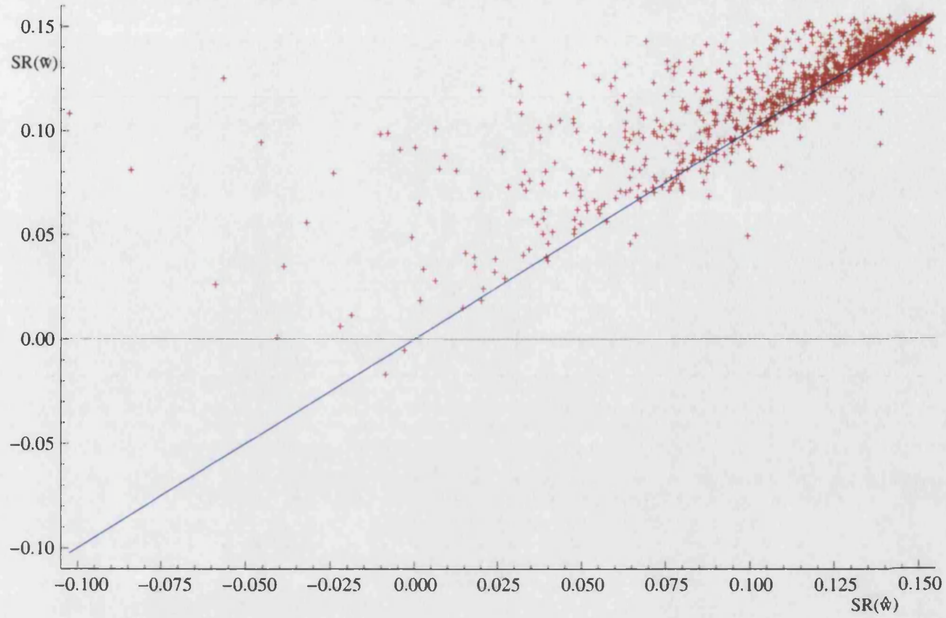

Figure 3.6: Sharpe Ratio of Tangency Portfolio under Two Estimators

estimator produces portfolios which have higher on average Sharpe Ratio and fewer disastrous results compared to the sample estimator.

Similar results are shown in Table 3.2.[7] On average the L_1 median estimator results in portfolios with higher Sharpe Ratio than the original sample estimator. Furthermore, 25% of the time, $\hat{\mathbf{w}}$ leads to portfolios with a Sharpe Ratio lower than 0.092. The corresponding quartile for $SR(\tilde{\mathbf{w}})$ is 0.107. Finally the most striking piece of evidence on the improvement achieved by $\tilde{\mathbf{w}}$ lies in the fact that out of 1000 simulated datasets, a significant 73% of the time $SR(\tilde{\mathbf{w}}) > SR(\hat{\mathbf{w}})$. This highlights the possible gains of this method.

Once again a visualisation of the improvement due to bootstrapping is shown in Figure 3.6.[8] It is evident that when the dataset is such that the sample estimator results in low (or even negative) "true" Sharpe Ratios, the bootstrap L_1 median acts as a protective shield against such unwanted results. At the same time, it performs equally well in datasets in which the sample tangency portfolio would achieve a high Sharpe Ratio.

[7]See also Table C.2.

[8]Corresponding plot is shown in Figure C.4.

Chapter 4

Portfolio Shrinkage Estimation

In Chapter 2, we explored a plethora of shrinkage estimators for the mean return and covariance matrix. In this section, in an attempt to improve portfolio estimation once again, we introduce two alternative shrinkage estimators. As before, stability is introduced by shrinkage towards desirable targets with the shrinkage intensity determined by taking into account the possible out-of-sample performance. The first estimator is general enough to include some of the already existing ones and therefore provides a ground for comparison. The second offers the flexibility of incorporating prior beliefs or knowledge regarding the portfolio composition.

4.1 Yet Another Shrinkage Estimator

We view the notion of shrinkage as a tool that enables us to create portfolios based on both observed data and prior beliefs. Before proceeding to introduce the first estimator we focus on some portfolios which are of particular importance in the finance literature. Apart from the global minimum variance portfolio and the tangency portfolio (both introduced in the first chapter), it is quite common to consider two additional portfolios: one is the equally weighted portfolio and the other is the mean proportional portfolio which assigns weights proportional to each asset's expected return, irrespective of the covariance matrix structure. It is therefore an advantage if, for some values of the shrinkage intensity parameters, the estimator includes the aforementioned portfolios.

One way of achieving this versatility for the portfolio weights vector is to consider the $\hat{\mathbf{w}}_{\alpha\beta}$ estimator.

Definition 2 *The $\hat{\mathbf{w}}_{\alpha\beta}$ estimator is defined as:*

$$\hat{\mathbf{w}}_{\alpha\beta} \propto \left[\alpha\,\mathbf{S} + (1-\alpha)\,\hat{\sigma}^2\mathbf{I}\right]^{-1}\left[\beta\,\bar{\mathbf{x}} + (1-\beta)\,\hat{\mu}\mathbf{1}\right] \quad \alpha, \beta \in [0,1] \tag{4.1}$$

subject to the full investment constraint $\mathbf{1}^T\hat{\mathbf{w}}_{\alpha\beta} = 1$. As usual, $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and covariance matrix respectively, while

$$\hat{\mu} = \frac{1}{N}\sum_{j=1}^{N}\bar{x}_j, \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{N}s_j^2$$

are the pooled scalar estimates of the average return and variance across all assets.

Shrinkage effectively transports us along two coordinate axes: one concerns the location vector and the other the dispersion matrix. We could use the sample mean and the unit vector as "endpoints" for the former axis and the sample covariance matrix and identity matrix for the latter, although this need not necessarily be the case. In fact, we will relax and further modify this restriction at a later stage to accommodate for additional information or investor beliefs.

As seen in Figure 4.1, this new coordinate system (α, β) maps all portfolios depending on the amount of shrinkage of their location vector and dispersion matrix away from the sample moments. It can also be verified that for suitable choices of α and β, this estimator can produce a variety of portfolios including the aforementioned desirable four. Therefore, the trivial choice of $\alpha = \beta = 0$ reduces to the equally weighted portfolio, denoted by Point Eq in Figure 4.1, while choosing $\alpha = 0$ and $\beta = 1$ produces the mean proportional portfolio (Point M). The global minimum variance portfolio (Point MV) is achieved by setting $\alpha = 1$ and $\beta = 0$ and the sample efficient tangency portfolio (Point T) is obtained by setting $\alpha = \beta = 1$.

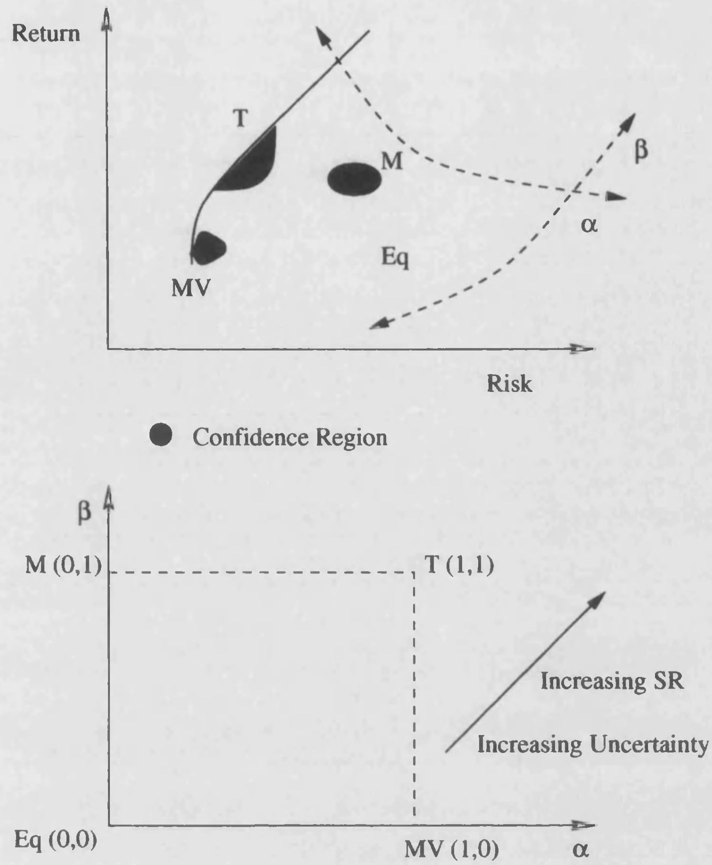

Figure 4.1: Mean Vector and Covariance Matrix Shrinkage Intensity

Apart from the portfolios of particular interest, this one-to-one mapping includes existing estimators such as the ones proposed by James and Stein (1961) and Ledoit (1995) and hence provides a common field onto which these estimators can be compared. In general any estimator which only shrinks the location vector while using the sample covariance matrix will always lie on the $\alpha = 1$ edge, while an estimator which applies shrinkage to the covariance matrix, while using the sample mean as an estimator of location will lie on the $\beta = 1$ edge.

4.2 Choosing Optimal α and β

As a result, the problem of selecting the "best" estimator is now reduced to selecting suitable α and β under some optimality criterion. We tackle this objective by turning our attention to the factors forcing us to apply shrinkage in the first place: although the sample efficient tangency portfolio maximises the (in-sample) Sharpe Ratio, it comes with a high degree of uncertainty as seen empirically in Figure 2.2 and illustratively in Figure 4.1. On the other hand, the global minimum variance portfolio is usually associated with lower Sharpe Ratio and at the same time lower uncertainty. A similar argument can be put forward for the mean proportional portfolio while the equally weighted portfolio comes with no uncertainty but with a potentially very low Sharpe Ratio. It can therefore be suggested that any optimality criterion for α and β should address the trade-off between higher Sharpe Ratio and lower uncertainty.

Each pair of values for α and β will result in portfolios with in-sample Sharpe Ratios which are at most equal to the sample efficient maximum Sharpe Ratio obtained from the tangency portfolio, i.e.

$$SR(\hat{\mathbf{w}}_{\alpha\beta}) = \frac{\hat{\mathbf{w}}_{\alpha\beta}^T \bar{\mathbf{x}}}{\sqrt{\hat{\mathbf{w}}_{\alpha\beta}^T \mathbf{S}\, \hat{\mathbf{w}}_{\alpha\beta}}} \leq SR(\hat{\mathbf{w}}_{11})$$

where $\hat{\mathbf{w}}_{11}$ denotes the estimated portfolio at $\alpha = \beta = 1$, i.e. the sample tangency portfolio. Therefore $SR(\hat{\mathbf{w}}_{11}) = \left(\bar{\mathbf{x}}^T \mathbf{S}^{-1} \bar{\mathbf{x}}\right)^{\frac{1}{2}}$ is the maximum attainable in-sample Sharpe Ratio. This prompts us to measure the "error" or "shortfall" in Sharpe Ratio performance in terms of the distance of $SR(\hat{\mathbf{w}}_{\alpha\beta})$ from the maximum $SR(\hat{\mathbf{w}}_{11})$ that is

$$\begin{aligned}\widehat{ShortFall}\left[SR(\hat{\mathbf{w}}_{\alpha\beta})\right] &= SR(\hat{\mathbf{w}}_{\alpha\beta}) - SR(\hat{\mathbf{w}}_{11}) \\ &= SR(\hat{\mathbf{w}}_{\alpha\beta}) - \left(\bar{\mathbf{x}}^T \mathbf{S}^{-1} \bar{\mathbf{x}}\right)^{\frac{1}{2}}\end{aligned}$$

The portfolio's shortfall due to the selection of appropriate pairs (α, β) should be weighted against the stability obtained by the shrinkage. Bootstrapping can be used to estimate the portfolio's Sharpe Ratio variability. We create B bootstrap resamples from the empirical distribution of the original data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ and denote each resample by $\left\{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_T^{(i)}\right\}$ for $i = 1, \ldots, B$. At each value of (α, β), the bootstrapped Sharpe Ratio is given by $SR\left(\hat{\mathbf{w}}_{\alpha\beta}^{(i)}\right)$ where $\hat{\mathbf{w}}_{\alpha\beta}^{(i)}$ is the $\hat{\mathbf{w}}_{\alpha\beta}$ estimator based on the ith resample. The portfolio's Sharpe Ratio variability is estimated by:

$$\widehat{Variability}\left[SR(\hat{\mathbf{w}}_{\alpha\beta})\right] = \frac{1}{B}\sum_{i=1}^{B}\left[SR\left(\hat{\mathbf{w}}_{\alpha\beta}^{(i)}\right) - SR(\hat{\mathbf{w}}_{11})\right]^2$$

Note that the variability is measured using the squared deviations from the *optimal* in-sample Sharpe Ratio rather than the bootstrapped mean Sharpe Ratio.

Schematically, the situation may be as depicted in Figure 4.2. Evaluating

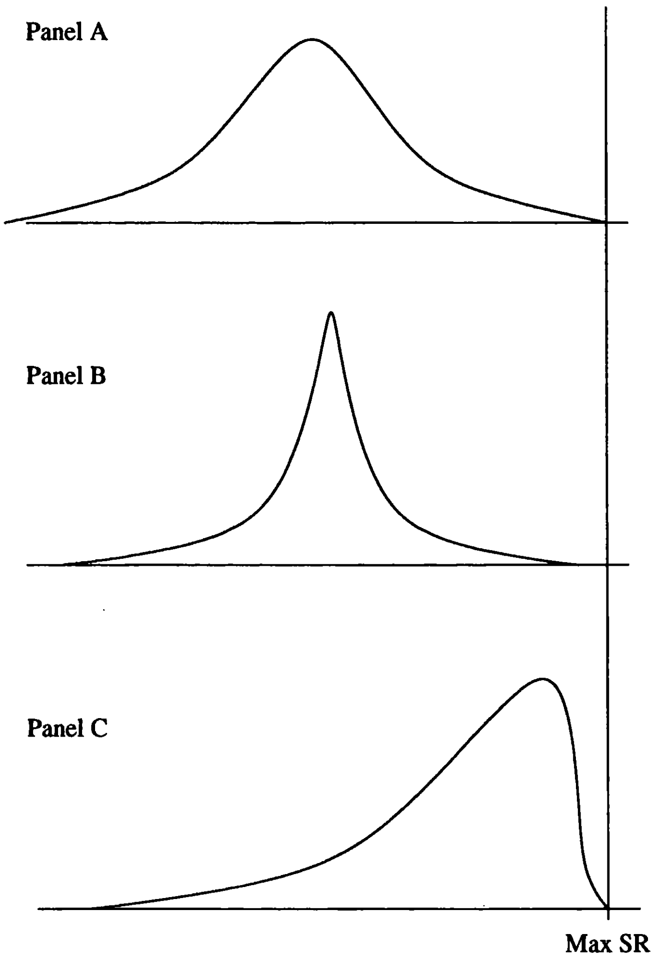

Figure 4.2: Possible Bootstrapped Distributions of in-sample Sharpe Ratio

the Sharpe Ratio over each bootstrapped resample for some (α, β) pairs may give rise to distributions exhibiting heavy tails as is the case in Panel A. This could be the case for estimators which are similar to the sample tangency portfolio and thus being governed by significant variability. On the other hand, applying shrinkage through the use of the (α, β) parameters may lead to alternative estimators which exhibit a smaller range of in-sample Sharpe Ratio values (Panel B). In fact, setting $\alpha = \beta = 0$ (i.e. using the equally weighted portfolio) would lead to a bootstrapped distribution with all of its mass on

$$SR(\hat{\mathbf{w}}_{00}) = \frac{\mathbf{1}^T\bar{\mathbf{x}}}{\mathbf{1}^T\mathbf{S}\,\mathbf{1}}.$$

Ideally, we would want to find such values of (α, β) which would produce high out-of-sample Sharpe Ratios. Therefore we would want the bootstrapped distribution of $SR(\hat{\mathbf{w}}_{\alpha\beta})$ to have most of its density as close to the maximum attainable Sharpe Ratio (Panel C), since, for each $\hat{\mathbf{w}}_{\alpha\beta}$ estimator based on a resample, $SR(\hat{\mathbf{w}}_{\alpha\beta})$ can be considered as its out-of-sample performance.

As a result of these considerations, one may choose the optimal (α, β) by minimising a loss function which depends on the possible shortfall and variability of the Sharpe Ratio i.e.

$$Penalty\,[\alpha,\beta] = \widehat{Variability}\,[SR(\hat{\mathbf{w}}_{\alpha\beta})] + \widehat{ShortFall}\,[SR(\hat{\mathbf{w}}_{\alpha\beta})]^2$$

subject to the constraints $\alpha, \beta \in [0, 1]$. Such a criterion would rank the estimator giving rise to the distribution in Panel C of Figure 4.2 as preferable to either of the other two estimators whose Sharpe Ratio bootstrapped distributions are depicted in Panels A and B.

The idea upon which the selected loss function is based stems from the Mean Squared Error measure. The main difference is that the variability term is measured from the maximum attainable Sharpe Ratio as opposed to the orthodox measure around the mean. One possible justification for this practice would be to view the chosen optimality criterion as a mixture of the realised squared deviation from the maximum attainable Sharpe Ratio and an average squared deviation of its out-of-sample performance.

Splitting the problem into two parts (Shortfall and Variability) highlights the Sharpe Ratio–uncertainty trade-off encountered in the (α, β) plane.

When $\alpha \approx 1$ and $\beta \approx 1$ the shortfall term will be very low since $\hat{\mathbf{w}}_{\alpha\beta}$ will be close to $\hat{\mathbf{w}}_{11}$. On the other hand, when $\alpha \approx 0$ and $\beta \approx 0$, the variability term will be small and in fact zero for the $\hat{\mathbf{w}}_{00}$ estimator (the equally weighted portfolio), since all portfolios, for all resamples will be identical. By minimising their sum we achieve a balance between the two.

An alternative possibility would be to consider, for each distribution in Figure 4.2, the $\gamma\%$ quantile and to choose the optimal (α, β) pair to maximise this. This would have the intuitive meaning of choosing an estimator by maximising a threshold value to be exceeded by its Sharpe Ratio $(1-\gamma)\%$ of the time. Typical values for γ could be 5, 10, or 25. This quantile approach takes into consideration the shape of the bootstrap distribution as well as its location and variability. Hence, it also accounts for the trade-off between Sharpe Ratio and uncertainty but in an indirect way.

4.3 Incorporating Prior Beliefs

The idea behind the $\hat{\mathbf{w}}_{\alpha\beta}$ estimator allows for a high degree of flexibility. We can change the shrinkage targets for either the dispersion matrix or the location vector to accommodate any possible prior information that may be available to an investor. We achieve this by modifying somewhat the optimization objective. The optimal shrinkage intensity parameters are chosen as before, by minimising a loss function which depends on a shortfall and a variability term of the sample equivalent of the objective function. Alternatively, the quantile approach may also be used.

We will illustrate here how we can choose alternative shrinkage targets. Let us suppose that prior beliefs[1] about the $N \times 1$ vector $\mathbf{w}$ are summarised by $\mathbf{w} \sim \mathcal{MN}(\mathbf{w}_0, \mathbf{V})$ under the constraints $\mathbf{1}^T\mathbf{w}_0 = 1$ and $\mathbf{V}\mathbf{1} = \mathbf{0}$, where $\mathbf{0}$ is a column vector of zeros. Note that the latter constraint is a consequence

[1] We use the term "prior beliefs" somewhat loosely since $\mathbf{w}$ is not really a random variable but a decision vector. Nevertheless, adopting this viewpoint allows for a framework in which to operate.

of $\mathbf{1}^T\mathbf{w} = 1$.

Such beliefs would indicate that the "true" portfolio would be centred around $\mathbf{w}_0$ with a transition structure $\mathbf{V}$. Alternatively, the transition matrix $\mathbf{V}$ may be viewed as quantifying the degree of reluctance in changing portfolio weights from their respective target $\mathbf{w}_0$. A diagonal element with small (large) value would indicate an investor's reluctance (willingness) to change that particular portfolio weight. On the other hand, off-diagonal elements would quantify an investor's behaviour in changing the portfolio structure and moving across assets.

It is therefore evident that the resulting shrinkage targets should be a function of the pre-determined parameters $\mathbf{w}_0$ and $\mathbf{V}$. One way of selecting these targets is derived from a parallel to the usual Bayesian framework: the prior information combined with the likelihood function produce the posterior distribution which summarises the information available for the parameters of interest. In this case the "parameter" that we are interested in is, in fact, the decision vector $\mathbf{w}$. Furthermore, the sample equivalent of the expected utility function in this framework has the role of the likelihood function in the Bayesian framework. Allowing for such relaxed conditions, consider for $\lambda_1 \geq 1, \lambda_2 \geq 0$ an estimator:

$$\begin{aligned} \hat{\mathbf{w}}_{\lambda_1\lambda_2} &= \arg\min_{\mathbf{w}} \{H(\mathbf{w})\} \\ H(\mathbf{w}) &= \frac{\lambda_1}{2}(\mathbf{1}^T\mathbf{S}^{-1}\bar{\mathbf{x}})\,\mathbf{w}^T\mathbf{S}\mathbf{w} - \mathbf{w}^T\bar{\mathbf{x}} + \frac{\lambda_2}{2}(\mathbf{w}-\mathbf{w}_0)^T\mathbf{V}^-(\mathbf{w}-\mathbf{w}_0) \end{aligned}$$

subject to the full investment constraint $\mathbf{1}^T\mathbf{w} = 1$. Note that the inverse of the matrix $\mathbf{V}$ does not exist since by construction it will be singular (due to the $\mathbf{V1} = \mathbf{0}$ constraint). Nevertheless, we may use the unique Moore-Penrose pseudoinverse $\mathbf{V}^-$ for its computation. It is easy to show that $\hat{\mathbf{w}}_{\lambda_1\lambda_2}$ is in a way an estimator based on shrinkage, if we rewrite the above optimization objective $H(\mathbf{w})$ as:

$$H(\mathbf{w}) \equiv \mathbf{w}^T\left(\frac{\lambda_1}{2}\mathbf{S} + \frac{\lambda_2}{2}\mathbf{V}^-\right)\mathbf{w} - \mathbf{w}^T\left(\bar{\mathbf{x}} + \lambda_2\mathbf{V}^-\mathbf{w}_0\right) + \frac{\lambda_2}{2}\mathbf{w}_0^T\mathbf{V}^-\mathbf{w}_0.$$

The shrinkage targets proposed by this analysis are the inverse transition matrix $\mathbf{V}^-$ and the transition adjusted weight vector $\mathbf{V}^-\mathbf{w}_0$.

The solution of this optimization problem can be shown to be

$$\begin{aligned}\hat{\mathbf{w}}_{\lambda_1\lambda_2} &= \frac{\mathbf{B}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{B}^{-1}\mathbf{1}} + \\ &+\mathbf{B}^{-1}\left(\bar{\mathbf{x}} - \frac{\mathbf{1}^T\mathbf{B}^{-1}\bar{\mathbf{x}}}{\mathbf{1}^T\mathbf{B}^{-1}\mathbf{1}}\mathbf{1}\right) + \\ &+\lambda_2\mathbf{B}^{-1}\left(\mathbf{V}^-\mathbf{w}_0 - \frac{\mathbf{1}^T\mathbf{B}^{-1}\mathbf{V}^-\mathbf{w}_0}{\mathbf{1}^T\mathbf{B}^{-1}\mathbf{1}}\mathbf{1}\right),\end{aligned}$$

where $\mathbf{B} = \lambda_1(\mathbf{1}^T\mathbf{S}^{-1}\bar{\mathbf{x}})\,\mathbf{S} + \lambda_2\mathbf{V}^-$.

Once again, this estimator covers a wide range of different portfolios. Setting ($\lambda_1 = 1, \lambda_2 = 0$) recovers the sample tangency portfolio, and it is for reasons of simplicity that the $(\mathbf{1}^T\mathbf{S}^{-1}\bar{\mathbf{x}})$ factor is included in the portfolio variance term. As λ_1 and λ_2 become larger, the estimator tends to other possibly desirable targets. More specifically, when λ_1 and λ_2 are large and $\lambda_1 \gg \lambda_2$ then we approximately have $\mathbf{B} \propto \mathbf{S}$, resulting approximately in $\hat{\mathbf{w}}_{\lambda_1\lambda_2} \propto \mathbf{S}^{-1}\mathbf{1}$, the global minimum variance portfolio. Alternatively, if $\lambda_1 \ll \lambda_2$ and large, the estimator $\hat{\mathbf{w}}_{\lambda_1\lambda_2} \to \mathbf{w}_0$ which is where the prior beliefs are centred around. Generally, a whole new variety of portfolios based on market information can be included in the $\hat{\mathbf{w}}_{\lambda_1\lambda_2}$ estimator.

4.4 Choosing Optimal λ_1 and λ_2

As before, optimal λ_1 and λ_2 are chosen in a way that takes into account the possible between-sample variability. The difference now is that rather than using the sample Sharpe Ratio as was the case for the $\hat{\mathbf{w}}_{\alpha\beta}$ estimator, we use the variability of the objective function $H(\mathbf{w})$ as a proxy for the uncertainty (and resulting instability) in the portfolio weights.

More specifically, the value of the objective function at the optimum will be $H(\hat{\mathbf{w}}_{\lambda_1\lambda_2})$ for different choices of λ_1 and λ_2. We are interested in the improvement achieved by incorporating market information, hence we compare the values of the λ-parameters with the benchmark position $\lambda_1 = 1, \lambda_2 = 0$ i.e. the parameters which would result in the sample tangency portfolio. In a similar manner as before, we are interested in the shortfall given by:

$$\widehat{ShortFall}\left[H(\hat{\mathbf{w}}_{\lambda_1\lambda_2})\right] = H(\hat{\mathbf{w}}_{\lambda_1\lambda_2}) - H(\hat{\mathbf{w}}_{10})$$

and variability

$$\widehat{Variability}\left[H(\hat{\mathbf{w}}_{\lambda_1\lambda_2})\right] = \frac{1}{B}\sum_{i=1}^{B}\left[H\left(\hat{\mathbf{w}}^{(i)}_{\lambda_1\lambda_2}\right) - H(\hat{\mathbf{w}}_{10})\right]^2$$

with $\hat{\mathbf{w}}^{(i)}_{\lambda_1\lambda_2}$ being the optimal weights derived from the ith resample. As a result, for the computation of $\hat{\mathbf{w}}^{(i)}_{\lambda_1\lambda_2}$ we would use the ith sample moments $\bar{\mathbf{x}}^{(i)}$ and $\mathbf{S}^{(i)}$, hence, $H\left(\hat{\mathbf{w}}^{(i)}_{\lambda_1\lambda_2}\right)$ would be the objective function value at these weights and $H(\hat{\mathbf{w}}_{10})$ the objective function at the sample efficient tangency portfolio. As before, we may choose optimal λ_1, λ_2 by

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \arg\min_{\lambda_1,\lambda_2}\left\{\widehat{Variability}\left[H(\hat{\mathbf{w}}_{\lambda_1\lambda_2})\right] + \widehat{ShortFall}\left[H(\hat{\mathbf{w}}_{\lambda_1\lambda_2})\right]^2\right\}.$$

Alternatively, the quantile approach may be used, although in such a case we need to consider the $(1-\gamma)\%$ quantile since the objective function $H(\hat{\mathbf{w}})$ is minimised (rather than maximised as was the Sharpe Ratio before). Both procedures will weigh the benefits of moving towards stable portfolios (in terms of low variability) against the costs of moving away from the sample efficient optimal portfolio (in terms of shortfall). The optimal solution achieves a balance between the two.

4.5 A Scenario with Market Information

The advantage provided by the $\hat{\mathbf{w}}_{\lambda_1\lambda_2}$ estimator lies in the fact that the parameters $\mathbf{w}_0$ and $\mathbf{V}$ can be chosen in such a way so that they reflect prior or market information available to the investor. Suppose an investor has some opinion about linear combinations of portfolio weights (such as the total investment on each sector of the market), and let $\mathbf{R}$ be a $K \times N$ matrix denoting these combinations. Since, $\mathbf{w} \sim \mathcal{MN}(\mathbf{w}_0, \mathbf{V})$ it follows that $\mathbf{Rw} \sim \mathcal{MN}(\mathbf{Rw}_0, \mathbf{RVR}^T)$ which summarises the information available to the investor. We focus here on the covariance matrix since choosing a target portfolio, $\mathbf{w}_0$ is the easier of the two tasks.

Full specification of the $N \times N$ covariance matrix $\mathbf{V}$, where

$$\mathbf{V} = \begin{pmatrix} v_1^2 & v_{12} & \cdots & v_{1N} \\ v_{12} & v_2^2 & \cdots & v_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1N} & v_{2N} & \cdots & v_N^2 \end{pmatrix},$$

requires $\frac{N}{2}(N+1)$ parameters. There are also N constraints due to the N equations $\mathbf{V1} = \mathbf{0}$. Therefore the free parameters are $\frac{N}{2}(N+1) - N = \frac{N}{2}(N-1)$. On the other hand, we can specify some of the elements of $\mathbf{V}$ and some additional elements of $\mathbf{RVR}^T$ and therefore infer the values of the remaining elements of $\mathbf{V}$. For example, it may be intuitively easier to assign prior values to the diagonal elements of $\mathbf{V}$ and $\mathbf{RVR}^T$ and by imposing some structure on the off-diagonal elements of $\mathbf{V}$, to deduce what the values of the remaining free parameters in $\mathbf{V}$ should be.

Consider for example the case where each of the stocks $j = 1, 2, \ldots, N$ belongs to one of K industries with $K < N$ (but usually $K \ll N$). To limit some of the parameters in $\mathbf{V}$ we make the following assumptions for each of the industries $g = 1, 2, \ldots, K$:

$$\left.\begin{array}{l} v_i^2 = v_j^2 = v_k^2 \\ v_{ij} = v_{ik} = v_{jk} \\ v_{il} = v_{jl} = v_{kl} \end{array}\right\} \forall \begin{array}{c} \{i, j, k\} \in \{S_g\} \\ l \notin \{S_g\} \end{array}$$

where $\{S_g\}$ denotes the set of stocks belonging to the gth industry. The first set of equations assumes that the portfolio weight variances is the same for all assets belonging to the same industries. The second set of equations restricts all pairwise within-industry covariances to a common value. Finally, we assume that a portfolio weight for an asset will have the same covariance with all other assets from a particular industry.

The number of distinct parameters in $\mathbf{V}$ is greatly reduced. There are now K diagonal elements to be specified (industry specific variances), K elements for the within-industry covariances and $\binom{K}{2}$ elements for the between-industry covariances. In total, there are $K+K+\frac{K}{2}(K-1)=K+\frac{K}{2}(K+1)$ parameters. The constraints due to $\mathbf{V1}=\mathbf{0}$ reduce the number of parameters by K so the total number of free parameters is $\frac{K}{2}(K+1)$. For the simple case $K=1$, the above structure allows only 1 free parameter and the resulting covariance matrix is:

$$\mathbf{\Omega}=\sigma^2\begin{pmatrix} 1 & -\frac{1}{N-1} & \cdots & -\frac{1}{N-1} \\ -\frac{1}{N-1} & 1 & \cdots & -\frac{1}{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{N-1} & -\frac{1}{N-1} & \cdots & 1 \end{pmatrix}$$

which assumes equal variances, equal covariances and satisfies $\mathbf{\Omega 1}=\mathbf{0}$. Intuitively this would mean that an investor is equally willing to change the portfolio weights of any asset for any other asset.

In general, this structure will result to the following covariance matrix:

$$\mathbf{V}=\begin{pmatrix} u_1^2 & u_{11} & \cdots & u_{11} & u_{12} & u_{12} & \cdots & u_{12} & \cdots & \cdots & u_{1K} \\ & u_1^2 & \cdots & u_{11} & u_{12} & u_{12} & \cdots & u_{12} & \cdots & \cdots & u_{1K} \\ & & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & u_1^2 & u_{12} & u_{12} & \cdots & u_{12} & \cdots & \cdots & u_{1K} \\ & & & & u_2^2 & u_{22} & \cdots & u_{22} & \cdots & \cdots & u_{2K} \\ & & & & & u_2^2 & \cdots & u_{22} & \cdots & \cdots & u_{2K} \\ & & & & & & \ddots & \vdots & \vdots & \vdots & \vdots \\ & & & & & & & u_2^2 & \cdots & \cdots & u_{2K} \\ & & & & & & & & \ddots & \vdots & \vdots \\ & & & & & & & & & \ddots & \vdots \\ & & & & & & & & & & u_K^2 \end{pmatrix}$$

where $u_g^2=Var(w_j)\ \forall j\in\{S_g\}$ denotes the industry-specific portfolio allocation variance, $u_{gg}=Cov(w_i,w_j)\ \forall\{i,j\}\in\{S_g\}$ denotes the within-industry weight covariance and $u_{gh}=Cov(w_i,w_j)\ \forall i\in\{S_g\},\ j\in\{S_h\}$ denotes the between-industry weight covariance.

Since we are working with assets belonging to industries, we can consider a combination matrix $\mathbf{R}$ with elements:

$$(\mathbf{R})_{gj} = \begin{cases} 1 & j \in \{S_g\} \\ 0 & j \notin \{S_g\} \end{cases}$$

for $j = 1, 2, \ldots, N$ and $g = 1, 2, \ldots, K$. Then let, $\mathbf{z} = \mathbf{Rw}$ be a $K \times 1$ vector of proportions invested in each industry and hence $\mathbf{1}^T\mathbf{z} = 1$. We can then denote by $\mathbf{\Sigma}_z = Var(\mathbf{z}) = \mathbf{RVR}^T$ the covariance matrix of the industry aggregated portfolio, which will also satisfy $\mathbf{\Sigma}_z\mathbf{1} = \mathbf{0}$, since the elements of $\mathbf{z}$ sum to 1. Thus the $\frac{K}{2}(K+1)$ parameters in:

$$\mathbf{\Sigma}_z = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1K} & \sigma_{2K} & \cdots & \sigma_K^2 \end{pmatrix}$$

are in fact reduced by K free parameters. If there were no restrictions on $\mathbf{\Sigma}_z$ this matrix would have uniquely determined $\mathbf{V}$. However, in order to completely determine $\mathbf{V}$, knowledge of $\mathbf{\Sigma}_z$ is not enough: we require knowledge of K additional elements of $\mathbf{V}$. Intuitively, it would make most sense to choose the diagonal elements of $\mathbf{V}$, and decide on the values of the off-diagonal elements through their relation with $\mathbf{\Sigma}_z$.

So far we have shown that, to completely specify the matrix $\mathbf{V}$, rather than determining all of its elements, we can make certain plausible assumptions to reduce the total number of free parameters. Furthermore, by using information that may be more intuitive (such as the industry portfolio weights' variances and covariances) to determine what the elements of $\mathbf{V}$ should be. More specifically, for the aforementioned example, let the number of stocks in $\{S_g\}$ be given by N_g for $g = 1, 2, \ldots, K$. Then, using notation defined above, we have

$$\begin{aligned} \sigma_g^2 &= N_g u_g^2 + N_g(N_g - 1)\, u_{gg} \\ \sigma_{gh} &= N_g N_h\, u_{gh} \end{aligned}$$

for $g \neq h$. Solving for u_{gg} and u_{gh} gives:

$$u_{gg} = \frac{\sigma_g^2 - N_g u_g^2}{N_g(N_g - 1)} \tag{4.2}$$

$$u_{gh} = \frac{\sigma_{gh}}{N_g N_h}. \tag{4.3}$$

We can also check that for any column in $\mathbf{V}$, the sum of the elements is given by:

$$u_g^2 + (N_g - 1)\, u_{gg} + \sum_{h \neq g} N_h u_{gh} =$$

$$= u_g^2 + \frac{\sigma_g^2 - N_g u_g^2}{N_g} + \sum_{h \neq g} \frac{\sigma_{gh}}{N_g} =$$

$$= \frac{1}{N_g} \left(\sigma_g^2 + \sum_{h \neq g} \sigma_{gh} \right) = 0$$

by our choice of elements in $\mathbf{\Sigma}_z$, which are the focus of attention of the next paragraph.

As it turns out, construction of an appropriate $\mathbf{\Sigma}_z$ is not difficult. Suppose that the initial prior $\mathbf{\Sigma}_z^*$ is only partly specified, and as a result $\mathbf{\Sigma}_z^* \mathbf{1} \neq \mathbf{0}$. Let $U = \mathbf{1}^T \mathbf{\Sigma}_z^* \mathbf{1}$ and $\mathbf{A} = \mathbf{\Sigma}_z^* \mathbf{1} / U$. Then, we can use:

$$\mathbf{\Sigma}_z = \alpha \left(\mathbf{\Sigma}_z^* - \mathbf{A}\mathbf{A}^T U \right)$$

which now satisfies $\mathbf{\Sigma}_z \mathbf{1} = \mathbf{0}$. The coefficient α can be chosen in such a way that $tr(\mathbf{\Sigma}_z) = tr(\mathbf{\Sigma}_z^*)$. In practice this means that we can decide *a priori* on some elements of the matrix $\mathbf{\Sigma}_z^*$ for which we have information (such as the diagonal elements) and adjust the remaining ones accordingly. Of course, one could argue that we could have applied the same treatment to the original prior matrix $\mathbf{V}$, but this two-step procedure allows us to impose some prior beliefs on intuitively more meaningful quantities such as the portfolio weights' variance for each industry.

4.6 A Simulated Example

We will be illustrating the two proposed estimators in the present section. We use the parameters $(\boldsymbol{\mu}, \mathbf{\Sigma})$ where:

$$\boldsymbol{\mu} = \frac{1}{12} (0.08, 0.06, 0.04, -0.01)^T$$

and

$$\boldsymbol{\Sigma} = \frac{1}{12}\,\mathrm{diag}\,(0.24^2, 0.18^2, 0.16^2, 0.15^2).$$

for the following comparison and simulation study.

4.6.1 The $\hat{\mathbf{w}}_{\alpha\beta}$ Estimator

We start by generating one sample of 120 observations from a multivariate normal distribution. The sample characteristics were calculated to be:

$$\bar{\mathbf{x}} = \frac{1}{12}(0.218, 0.113, 0.003, -0.051)^T$$

and

$$\mathbf{S} = \frac{1}{12}\begin{pmatrix} 0.060 & 0.002 & 0.001 & -0.003 \\ 0.002 & 0.029 & -0.001 & 0.002 \\ 0.001 & -0.001 & 0.024 & -0.001 \\ -0.003 & 0.002 & -0.001 & 0.020 \end{pmatrix}$$

which resulted in estimated minimum variance and tangency portfolios given by

$$\begin{aligned} \hat{\mathbf{w}}_{GMV} &= (0.12, 0.22, 0.27, 0.38)^T \\ \hat{\mathbf{w}}_T &= (0.30, -0.01, 0.19, 0.52)^T. \end{aligned}$$

Firstly we show how the $\hat{\mathbf{w}}_{\alpha\beta}$ estimator would be obtained and then we show the results of a simulation study.

A total of $B = 500$ bootstrap resamples were generated, based on the original dataset. Using the procedure of calculating optimal parameters $\hat{\alpha}$ and $\hat{\beta}$ by minimising the decomposed (shortfall and variability) penalty of the sample Sharpe Ratio as explained in this chapter, we find

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 0.72 \\ 0.36 \end{pmatrix}.$$

From this we conclude that better results will be obtained if more shrinkage is applied to the location vector rather than the dispersion matrix (since no shrinkage is equivalent to $\alpha = \beta = 1$).

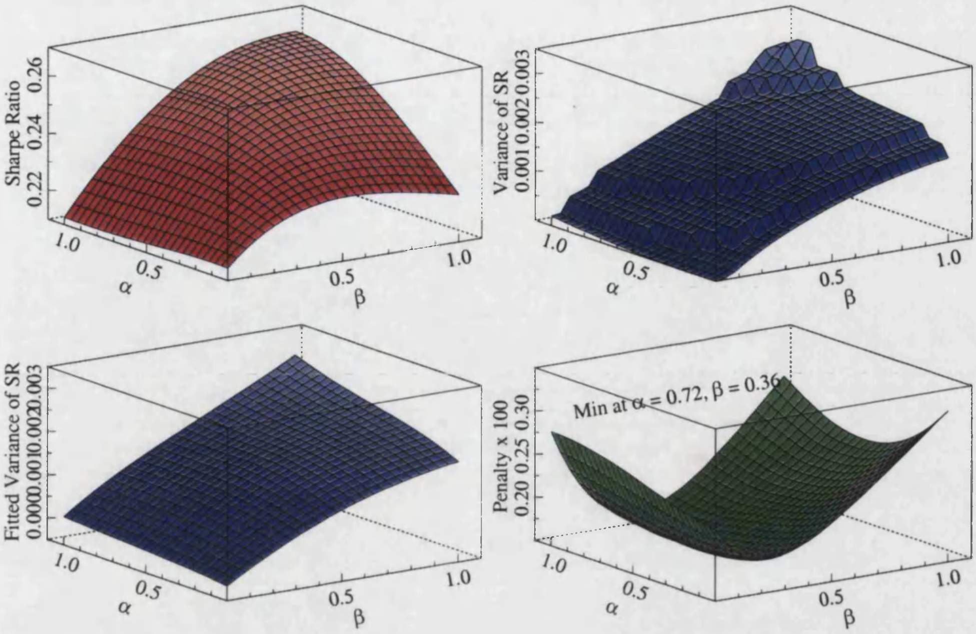

Figure 4.3: Minimising the Penalty of the Sample Tangency Portfolio Sharpe Ratio.

The four panels in Figure 4.3 illustrate the trade-off between the in-sample Sharpe Ratio and the uncertainty. The upper left hand panel shows the in-sample Sharpe Ratio for a grid of values of α and β. As α and β increase, thus the portfolio moves towards the tangency portfolio, the sample Sharpe Ratio increases as well. However, as can be seen by the upper right hand panel of Figure 4.3 the estimated variability of the sample Sharpe Ratio, measured over the 500 bootstrap resamples increases in the same manner. A smoothed version of this graph is given in the lower left hand panel. The equal weighted portfolio (obtained by setting $\alpha = \beta = 0$) has no uncertainty since for every resample it will always be the same irrespective of the bootstrap sample moments. As the portfolio approaches the tangency portfolio, uncertainty increases dramatically.

The bottom right hand panel shows the overall penalty. It is equal to the sum of the squared shortfall and the estimated variability of the sample Sharpe Ratio. As mentioned before the shortfall is measured as the distance from the optimal in-sample Sharpe Ratio, i.e. as the vertical distance of any point on the surface in the upper left hand panel from the Sharpe Ratio at $\alpha = \beta = 1$.

It is now obvious that by minimising the penalty term, we achieve a balance on the trade-off between the accuracy of the estimator and its out-of-sample performance. Any possible increase in the sample Sharpe Ratio is counteracted by a possible increase in the instability of the estimator.

We now turn our attention to a simulation study designed to show how the $\hat{\mathbf{w}}_{\alpha\beta}$ estimator would fare in comparison with the sample moments tangency portfolio estimator. In a similar manner as before, we generate 1000 samples from the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For each sample we calculate the $\hat{\mathbf{w}}_T$ and $\hat{\mathbf{w}}_{\alpha\beta}$ estimators and compare them in the basis of the "true" Sharpe Ratio, i.e. the Sharpe Ratio that would have been obtained under the population parameters.

Some summary statistics on the distribution of the Sharpe Ratio under the two estimators are given in Table 4.1.[2] It is again obvious that the

	$SR(\hat{\mathbf{w}}_T)$	$SR(\hat{\mathbf{w}}_{\alpha\beta})$
Min	-0.102	-0.018
Q1	0.092	0.120
Mean	0.110	0.127
Median	0.119	0.132
Q2	0.137	0.138
Max	0.155	0.155
Res	277	723

Table 4.1: Summary Statistics for $SR(\hat{\mathbf{w}}_{\alpha\beta})$

$\hat{\mathbf{w}}_{\alpha\beta}$ estimator produces on average portfolios with higher Sharpe Ratio, as

[2]One may notice that the same bootstrap resamples were used as in the simulated example of the previous chapter, hence the summary statistics under the $\hat{\mathbf{w}}_T$ estimator are the same. They are reproduced here for comparison.

indicated by the mean Sharpe Ratio. The quartiles and the median are also higher in the case of the $\hat{\mathbf{w}}_{\alpha\beta}$ and furthermore, out of the 1000 simulated samples, the shrinkage estimator was better compared to the sample tangency portfolio in 723 datasets.

Figure 4.4 conveys even more information. As in the previous chapter, it

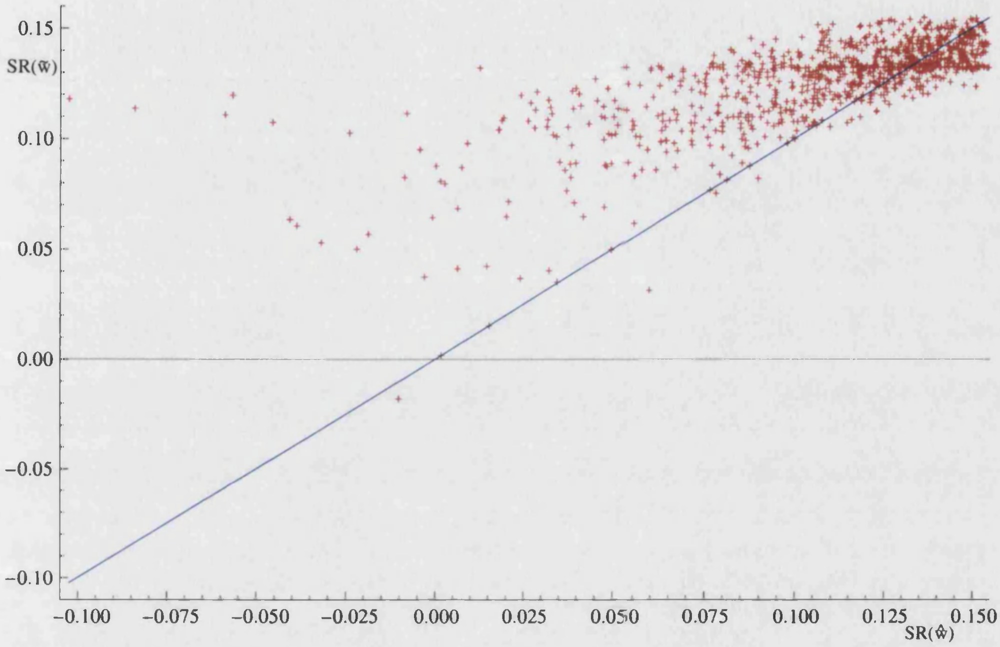

Figure 4.4: Scatter Plot of Sharpe Ratio under $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}_{\alpha\beta}$.

plots the 1000 pairs of true Sharpe Ratio for the two estimators. The $\hat{\mathbf{w}}_{\alpha\beta}$ estimator is shown in the vertical axis against the sample tangency portfolio which is the benchmark on the horizontal axis. One point of significance is that shrinkage produces portfolios which will continue performing well even if the observed dataset gives rise to unstable sample tangency portfolios.

This is the case with the portfolios on the left hand side of the graph (i.e. with low $SR(\hat{\mathbf{w}}_T)$). Choosing the shrinkage intensity through bootstrap resampling takes into account the possible uncertainty and as a result leads to a portfolio with better out-of-sample characteristics. The 'true' Sharpe Ratio exhibited by the $\hat{\mathbf{w}}_{\alpha\beta}$ estimator is considerably higher in this area of the graph.

On the other hand, when the resample moments are similar to the sample ones, the sample tangency portfolio would give rise to high Sharpe Ratios. This corresponds to points on the right hand side of the graph. In this case the shrinkage estimator fares equally well with $\hat{\mathbf{w}}_T$ with some points above and some below the 45-degree, equivalence, line.

Overall the new estimator offers a considerable improvement on the existing sample moments estimator as this simulation has shown.

4.6.2 The $\hat{\mathbf{w}}_{\lambda_1 \lambda_2}$ Estimator

We demonstrate here the steps followed in choosing a suitable transition matrix to reflect prior beliefs or market information. For this example, we use the same generated sample with moments $\bar{\mathbf{x}}$ and $\mathbf{S}$ as the first simulation in this section.

Using notation introduced earlier in this chapter, we assume that each of the available $N = 4$ assets, belongs to either of two ($K = 2$) industries. Let us further assume that the first two assets belong to the first industry whereas the remaining two belong to the second one. Suppose also that an investor targets a portfolio with composition $\mathbf{w}_0 = (0.1, 0.2, 0.3, 0.4)^T$ and that, although willing to change the allocation among the assets of an industry, (s)he prefers to keep the industry portfolio allocation at approximately $\mathbf{z} = (0.3, 0.7)^T$. Under these circumstances the investor's willingness (or lack of) to change any asset (with $v_1^2 = v_2^2$ and $v_3^2 = v_4^2$) from either industry could be assumed to be, for example, $u_1^2 = 1$ and $u_2^2 = 0.1$, whereas the reluctance of changing the industry portfolio to be $\sigma_1^2 = \sigma_2^2 = 0.01$. It is important here to note, that it is the *relative* magnitude of these values that is important rather than the actual values, given that any proportionality coefficient will be unimportant due to the parameter λ_2.

Based on these assumptions, and following notation in this chapter, we have so far

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} 1.0 & & & \\ & 1.0 & & \\ & & 0.1 & \\ & & & 0.1 \end{pmatrix}$$

whereas $\mathbf{\Sigma}_z = \mathbf{RVR}^T$ must be given by:

$$\mathbf{\Sigma}_z = \begin{pmatrix} 0.01 & -0.01 \\ -0.01 & 0.01 \end{pmatrix}$$

so that it satisfies $\mathbf{\Sigma}_z \mathbf{1} = \mathbf{0}$. Using equations (4.2) and (4.3) with $N_1 = N_2 = 2$ and $\sigma_{12} = \sigma_{21} = -0.01$ leads to full specification of the transition matrix $\mathbf{V}$ i.e.

$$\mathbf{V} = \begin{pmatrix} 1 & -0.995 & -0.0025 & -0.0025 \\ -0.995 & 1 & -0.0025 & -0.0025 \\ -0.0025 & -0.0025 & 0.1 & -0.095 \\ -0.0025 & -0.0025 & -0.095 & 0.1 \end{pmatrix}.$$

Equipped with the transition matrix $\mathbf{V}$ and the target vector $\mathbf{w}_0$ we proceed with the optimization. For this example, the optimization parameters were calculated to be

$$\begin{pmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{pmatrix} = \begin{pmatrix} 6.696 \\ 0.014 \end{pmatrix},$$

and the estimated portfolio weights

$$\hat{\mathbf{w}}_{\lambda_1\lambda_2} = \begin{pmatrix} 0.14 \\ 0.16 \\ 0.28 \\ 0.42 \end{pmatrix}.$$

It is evident that, although each portfolio weight in $\hat{\mathbf{w}}_{\lambda_1\lambda_2}$ is different than the corresponding element in $\mathbf{w}_0$, the proportion invested in each industry remained approximately equal ($0.14 + 0.16 = 0.3$, $0.28 + 0.42 = 0.7$). Moreover, as seen from the transition matrix $\mathbf{V}$ investing more in asset 1 would be financed by selling asset 2 due to the negative off-diagonal element

u_{11} (denoting the within-industry "covariance"). On the other hand transactions between industries would be rare due to the (relatively) small values $\sigma_1^2 = \sigma_2^2 = 0.01$. Finally, the investor would be more willing to trade assets from the first industry rather than the second since $u_1^2 > u_2^2$.

This small example demonstrates the significant capabilities of the $\hat{\mathbf{w}}_{\lambda_1\lambda_2}$ estimator. The transition matrix can be chosen in such way to reflect more complicated market information and investor preferences. As a result it may prove to be invaluable in real-life scenarios when additional information should be accounted for.

4.7 A Further Note on Shrinkage

We briefly diverge in this section to note an alternative way of accounting for uncertainty in portfolio selection. As has been argued, some shrinkage estimators can be recovered by including penalty terms in the optimization objective and solving the modified problem. Such a penalty term is usually linked with the uncertainty governing regions of the solution space.

Earlier in this chapter we have argued how bootstrap resamples can be used to assess the out-of-sample performance, by uncovering the sensitivity of different estimators to changes in the data. An estimator $\hat{\mathbf{w}}_1$ is a possibly non-linear function of the data, that is $\hat{\mathbf{w}}_1 = f(\mathbf{x}_1, \ldots, \mathbf{x}_t)$. Bootstrap resamples are perturbations of the original observed sample, whereas the resulting bootstrap distributions of an estimator show its responsiveness to these perturbations. A different way of viewing this responsiveness is through derivatives.

Effectively we are interested in how $\hat{\mathbf{w}}_1$ changes if the estimator was based on a marginally different sample. For example, one could consider an estimator computed as $\hat{\mathbf{w}}_2 = f(\mathbf{x}_1, \ldots, \mathbf{x}_T + \epsilon)$, for small ϵ and observe the changes from the original estimator $\hat{\mathbf{w}}_1$. Alternatively, we could compute $\hat{\mathbf{w}}_3 = f(\mathbf{x}_1 + \epsilon_1, \ldots, \mathbf{x}_T + \epsilon_T)$ for suitably small $\epsilon_1, \ldots, \epsilon_T$.

Depending on the nature of the original estimator, the changes in the data could have a significant, moderate, or negligible effect. On the other hand, it could be the case that they have no effect at all. The sample tangency portfolio, for example, would be highly sensitive to such changes, whereas their effect will be considerably less on the sample minimum variance portfolio. The equally weighted portfolio $\hat{\mathbf{w}} \propto \mathbf{1}$, on the other hand, would always remain the same, irrespective of the observed sample.

These considerations point towards using a suitable penalty term based on the first derivatives of the elements of each estimator with respect to the data. Portfolio estimators are $N \times 1$ vectors and are computed as functions of T multivariate data points. Hence, one could compute the T matrices of first derivatives (each of dimension $N \times N$) of the effect of a marginal change in each data point on the estimator and find the average. The (i, j) element of the average matrix of first derivatives would reflect the average effect of the ith element of each datapoint on the jth element of the estimator. An instability measure can thus be constructed. The optimal portfolio would then be selected to achieve a balance between in-sample performance and stability. As before, portfolios which have a high in-sample performance will be more likely to have a high degree of instability and thus would be penalised.

Chapter 5

Portfolio Selection in Practice

In this chapter, we will be viewing the portfolio optimization problem from a practitioner's perspective. Apart from expected returns and covariances, a rational investor will take into account portfolio constraints and, most importantly, transaction costs before making any decisions. Once these parameters have been considered, an optimal portfolio will be chosen. This can also be viewed as the first step towards sequential optimization since the starting portfolio influences the decision for the optimal final holdings.

5.1 Utility Theory with Transaction Costs

Let us start from the utility framework introduced in the first chapter. Assume that an investor's current portfolio holdings and wealth are given by $\mathbf{w}_{t-1}$ and M_{t-1} respectively. Furthermore assume that, the cost incurred for any transaction is proportional to the amount of transacted wealth, irrespective of whether the investor is buying or selling. Let the transaction cost rate be λ, meaning that λ units of currency are to be paid for each unit of transacted wealth. As a result, the investor's next period wealth will be given by:

$$M_t = M_{t-1}\left(1 - \lambda\,\mathbf{1}^T\mathbf{d}\right)\,\mathbf{w}^T\left(\mathbf{1} + \mathbf{x}_t\right).$$

where $\mathbf{d} = \{d_j : j = 1, \ldots, N\}$ and $d_j = |w_j - w_{j,t-1}|$ denotes the vector of absolute differences between the elements of $\mathbf{w}$ and $\mathbf{w}_{t-1}$. The after-cost portfolio return will therefore be:

$$\begin{aligned}
R_t &= \frac{M_t - M_{t-1}}{M_{t-1}} \\
&= \frac{M_{t-1}\left[\left(1 - \lambda\, \mathbf{1}^T\mathbf{d}\right)\,\left(1 + \mathbf{w}^T\mathbf{x}_t\right) - 1\right]}{M_{t-1}} \\
&= \left(1 - \lambda\, \mathbf{1}^T\mathbf{d}\right)\mathbf{w}^T\mathbf{x}_t - \lambda\, \mathbf{1}^T\,\mathbf{d} \\
&\approx \mathbf{w}^T\mathbf{x}_t - \lambda\, \mathbf{1}^T\mathbf{d}.
\end{aligned}$$

This approximation essentially omits the term $-\lambda\,(\mathbf{1}^T\mathbf{d})(\mathbf{w}^T\mathbf{x}_t)$ which corresponds to the effect of *not* investing the transaction costs, on the portfolio return. Its omission simplifies some of the computational issues and is justified since the omitted term is very small compared to the portfolio return. In real terms, this would be the portfolio return of investors who charge any transaction costs $\lambda\,\mathbf{1}^T\mathbf{d}$ to a secondary bank account. For our objectives, the approximation is adequate.

Using the negative exponential utility function $U(R_t) = 1 - \exp\{-\kappa R_t\}$ or more specifically,

$$U(R_t) = 1 - \exp\,\left\{-\kappa\left(\mathbf{w}^T\mathbf{x}_t - \lambda\,\mathbf{1}^T\mathbf{d}\right)\right\}$$

we get:

$$E\left[U(R_t)\right] = 1 - \exp\left\{\kappa\left(\frac{\kappa}{2}\,\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} - \mathbf{w}^T\boldsymbol{\mu} + \lambda\,\mathbf{1}^T\mathbf{d}\right)\right\}.$$

Under the utility framework, rational investors maximize their expected utility which is equivalent to choosing optimal portfolio weights $\mathbf{w}_L$ such that:

$$\mathbf{w}_L = \arg\min_{\mathbf{w}}\left\{\frac{\kappa}{2}\,\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} - \mathbf{w}^T\boldsymbol{\mu} + \lambda\,\mathbf{1}^T\mathbf{d}\right\} \tag{5.1}$$

subject to the full investment constraint $\mathbf{1}^T\mathbf{w} = 1$. Since the true population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, investors will use estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ in their place.

5.2 The Lasso - A Related Concept

Objective (5.1) is a non-linear optimization problem. However, it is at this point that we exploit a striking similarity between objective (5.1) and the Lasso estimator, proposed by Tibshirani (1996). Consider the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is the design matrix of explanatory variables, $\mathbf{y}$ is the response variable, $\boldsymbol{\beta}$ is the vector of model coefficients and $\boldsymbol{\epsilon}$ is the error vector. Rather than using the ordinary least squares estimator, Tibshirani (1996) considered the vector of coefficients $\mathbf{b}_L$ such that

$$\mathbf{b}_L = \arg\min_{\mathbf{b}} \left\{ (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) + \lambda \sum_j |b_j|^\gamma \right\}, \quad \lambda \geq 0, \gamma > 0. \quad (5.2)$$

Various values of γ and λ give rise to different known estimators in the statistical literature. The usual minimization of squared residuals is obtained by setting $\lambda = 0$. When $\gamma = 1$, we obtain the "least absolute shrinkage and selection operator", namely the *Lasso* estimator. Finally, $\gamma = 2$ leads to the ridge regression setting.

The reasoning behind the constraints imposed on the parameter vector $\boldsymbol{\beta}$ is twofold. On the one hand, the new estimator may have a significantly reduced variance, albeit at the expense of some bias compared to the ordinary least squares estimator. On the other hand, as argued by Tibshirani (1996) and later by Fu (1998), an important advantage of the Lasso estimator is that it shrinks only some of the coefficients towards zero, while setting others exactly to zero. This is in contrast to ridge regression, whereby shrinkage towards zero is applied for all coefficients. In our case, setting a coefficient equal to exactly zero will be equivalent to setting the portfolio weight of one particular asset, w_j, to its respective value in the previous period, $w_{j,t-1}$, avoiding therefore any transaction costs on that asset.

Having recognised the analogy between objectives (5.1) and (5.2), we propose the employment of computation methods for the lasso estimator to the portfolio optimization problem with transaction costs.

5.3 The Modified Shooting Algorithm

Fu (1998) describes a "shooting algorithm" for the case of $\gamma = 1$ in objective (5.2). Here, we modify this algorithm to optimise objective (5.1) with $\bar{\mathbf{x}}$ and $\mathbf{S}$ replacing their population counterparts, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. We will include an additional variable in the algorithm. This, first alteration, is necessary to enforce the full investment constraint.

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ be the $T \times N$ matrix of excess returns. Then

$$\bar{\mathbf{x}} = \frac{\mathbf{X}^T\mathbf{1}}{T}, \quad \mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{T} - \bar{\mathbf{x}}\bar{\mathbf{x}}^T$$

and the sample equivalent of the objective function in (5.1), including the Lagrange multiplier now becomes

$$H(\mathbf{w}, \phi) = \frac{\kappa}{2}\mathbf{w}^T\mathbf{S}\mathbf{w} - \mathbf{w}^T\bar{\mathbf{x}} + \lambda\,\mathbf{1}^T\mathbf{d} + \phi\,(\mathbf{1}^T\mathbf{w} - 1).$$

Differentiation with respect to w_j for $j = 1, \ldots, N$ yields N equations:

$$\frac{\partial H}{\partial w_j} = m_j w_j + c_j + \lambda\,\mathrm{sgn}\,(d_j)$$

where

$$m_j = \kappa\left(\frac{\mathbf{x}_j^T\mathbf{x}_j}{T} - \bar{x}_j^2\right),$$

$$c_j = \phi - \bar{x}_j + \kappa\sum_{i \neq j}\left(\frac{\mathbf{x}_j^T\mathbf{x}_i}{T} - \bar{x}_j\bar{x}_i\right)w_i$$

and

$$\mathrm{sgn}\,(d_j) = \begin{cases} 1 & \text{if } d_j > 0 \\ 0 & \text{if } d_j = 0 \\ -1 & \text{if } d_j < 0 \end{cases}.$$

In the same manner as in Fu (1998), there will be at most one solution for each one of the N equations,

$$m_j w_j + c_j + \lambda\,\mathrm{sgn}\,(d_j) = 0, \quad j = 1, \ldots, N$$

or

$$m_j w_j + c_j = -\lambda \operatorname{sgn}(d_j) \tag{5.3}$$

if and only if $m_j > 0, \forall j$. Since $m_j = \kappa\, s_{x_j}^2$ where $s_{x_j}^2$ is the variance of the jth asset return and $\kappa > 0$, these conditions are met.

We illustrate the solution of this problem in Figure 5.1. The first panel

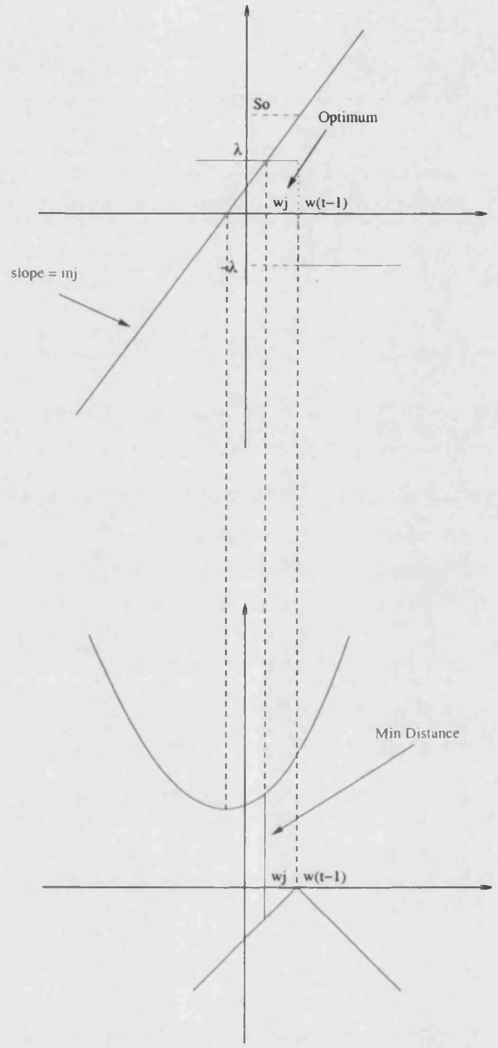

Figure 5.1: The Shooting Method

shows both the marginal gain (blue line) and marginal loss (red line) in the optimization objective (5.1), for the jth element of $\mathbf{w}$. The former is the left hand side of equation (5.3) whereas the latter is the right hand side. The marginal gain is an upward sloping line (since $m_j > 0$) which will either intersect the marginal cost at $\pm\lambda$ or not intersect the marginal cost line at all. In this case, we take the solution to lie at the original point $w_{j,t-1}$. Furthermore, the w_j co-ordinate of the intersection of the marginal gain and loss is a non-increasing function of the Lagrange multiplier ϕ. Evidently as ϕ increases, the point of intersection moves towards the left, unless the solution is at $w_{j,t-1}$. In this case, the w_j co-ordinate remains constant.

The second panel plots the actual objective, $H(\mathbf{w}, \phi)$, as a function of w_j. Once again, $H(\mathbf{w}, \phi)$ is decomposed into the gain function (blue curve) and the negative loss function (red line) which correspond to their marginal equivalents in the first panel. We are attempting to minimise the sample equivalent of objective (5.1) which is the vertical distance between the curve and the discontinuous line. The solution is found at the point where the marginal gain is equal to the marginal loss (i.e. where the tangent of the curve has the same gradient as the line), or at the point of discontinuity.

Before proceeding with the optimization algorithm, we focus on the Lagrange multiplier ϕ. Its effect is to enforce the full investment constraint $\mathbf{1}^T\mathbf{w} = 1$ which will be satisfied at a particular, unknown value of ϕ. As a result, apart from the optimal w_j's, the algorithm needs to optimise the Lagrange multiplier as well. Towards this objective, we will use an iteration procedure based on the fact that the quantity $\sum w_j$ is a non-increasing function of ϕ.

The computation of $\mathbf{w}$ and ϕ is simple. We summarise the modified shooting algorithm as follows. First of all, for a given value of ϕ, define $S_j(w_j, \mathbf{w}^{-j}, \phi) = m_j w_j + c_j$ where $\mathbf{w}^{-j}$ denotes the remaining elements of $\mathbf{w}$ after omitting w_j, so that $\mathbf{w} = (w_j, \mathbf{w}^{-j})^T$. Then:

Modified Shooting Algorithm

1. Start with a value ϕ_0 for the Lagrange multiplier.

2. Start the iteration with an estimator $\hat{\mathbf{w}}_0$

3. At iteration i, for each $j = 1, \ldots, N$ let $S_0 = S_j(w_{j,t-1}, \hat{\mathbf{w}}_{i-1}^{-j}, \phi_{k-1})$ and set

$$\hat{w}_j = \begin{cases} w_{j,t-1} + \frac{\lambda - S_0}{m_j} & \text{if } S_0 > \lambda \\ w_{j,t-1} + \frac{-\lambda - S_0}{m_j} & \text{if } S_0 < -\lambda \\ w_{j,t-1} & \text{if } |S_0| \leq \lambda \end{cases}$$

Form a new estimator $\hat{\mathbf{w}}_i = (\hat{w}_1, \ldots, \hat{w}_N)^T$ after all $\hat{w}_j$'s have been updated.

4. Repeat Step 3 until convergence of $\hat{\mathbf{w}}_i$.

5. Update ϕ by using $\phi_k = \phi_{k-1} + 2^{1-k}\, \alpha \operatorname{sgn}(\mathbf{1}^T \hat{\mathbf{w}}_i - 1)$

6. Repeat Steps 3-5 until convergence of ϕ_k.

The Modified Shooting Algorithm starts with an arbitrary choice ϕ_0. After optimization of the objective function $H(\mathbf{w}, \phi_0)$, the vector $\hat{\mathbf{w}}$ will not necessarily satisfy the constraint $\mathbf{1}^T\hat{\mathbf{w}} = 1$ and thus the Lagrange multiplier will need to be updated. As we have argued, the quantity $\sum \hat{w}_j$ will be a non-increasing function of ϕ and hence if $\mathbf{1}^T\hat{\mathbf{w}} > 1$, the optimal value ϕ^* must satisfy $\phi^* > \phi_0$. Under a similar argument, if $\mathbf{1}^T\hat{\mathbf{w}} < 1$, we must have $\phi^* < \phi_0$. The iteration procedure for optimising ϕ starts with a jump of α units towards the optimal direction. Assuming that at the next step, $\phi_0 < \phi^* < \phi_1$ or $\phi_1 < \phi^* < \phi_0$ depending on the optimal direction, the updating rule $\phi_k = \phi_{k-1} + 2^{1-k}\, \alpha \operatorname{sgn}(\mathbf{1}^T\hat{\mathbf{w}} - 1)$ will bisect the range of values of ϕ until $\phi = \phi^*$ and thus $\mathbf{1}^T\hat{\mathbf{w}} = 1$ within a degree of specified precision. Optimization of $H(\mathbf{w}, \phi_k)$ will be required at each step but convergence is achieved without excessive computational time.

5.4 A More General Scenario

The powerful nature of this procedure is that, assuming that the conditions $m_j > 0$ hold, much more complicated cost structures can be incorporated in the optimization, including portfolio constraints. In this section, through the portfolio selection framework, we propose such extensions which generalise the algorithm. This has the added advantage of rendering the method suitable to additional applications.

For example, consider the scenario depicted in Figure 5.2. First of all we

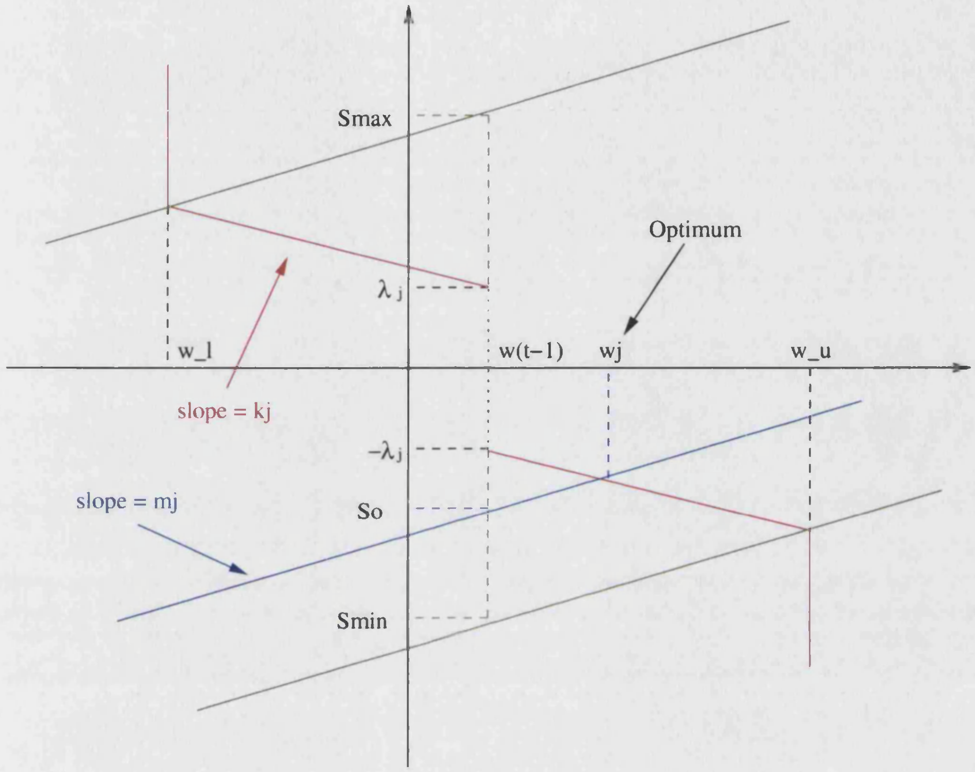

Figure 5.2: The Shooting Method for Increasing Transaction Costs and Portfolio Constraints

can assign a different transaction cost rate, λ_j, for each asset. Investors may find that some assets are more expensive to trade because of their (lack of) liquidity. Furthermore, the transaction cost rate λ_j may increase as the portfolio weights move further away from the original portfolio. This describes the situation faced by a large company: small adjustments are subject to the usual costs, whereas significant changes may affect the whole market and hence may incur a higher (marginal) cost. On the other hand, a company or an individual may be subject to certain constraints. An individual will not have a portfolio with negative weights, whereas companies usually constrain the portfolio weights within certain limits.

More specifically, denote the vector of absolute differences $\mathbf{d}$ as before with $d_j = |w_j - w_{j,t-1}|$ for $j = 1, \ldots, N$. So far we have been focusing on the case that the transaction costs are equal to $f(\mathbf{d}) = \lambda \mathbf{1}^T \mathbf{d}$, which has constant marginal costs $\frac{\partial f}{\partial d_j} = \lambda \operatorname{sgn}(d_j)$. We can extend the transaction cost structure to include both different base costs and increasing marginal costs for each asset, as in the case of:

$$f(\mathbf{d}) = \boldsymbol{\lambda}^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T \mathbf{K} \mathbf{d},$$
$$\mathbf{K} = \operatorname{diag}(k_1, k_2, \ldots, k_N)$$

with

$$\frac{\partial f}{\partial d_j} = \operatorname{sgn}(d_j)\left(\lambda_j + k_j d_j\right).$$

Portfolio constraints are in effect boundaries beyond which transaction costs (or any other penalty term) would become infinite. This again is easily incorporated in this method as shown in Figure 5.2. Portfolio weight w_j is penalised by λ_j units for any movement away from the starting position $w_{j,t-1}$, and an additional k_j units per unit of movement. Once it reaches either of boundaries w_j^l or w_j^u the marginal cost becomes infinite and the optimal weight would be at the boundary. The portfolio constraint could also be reflected in the lower panel of Figure 5.1 by vertical lines in the negative loss function.

The shooting algorithm is once again modified in a similar manner. The solution region of w_j is specified by the value S_0 in Figure 5.2: if $S_0 \geq S_{max}$ then the solution lies on the lower boundary for w_j, that is w_j^l. On the other hand, if $S_0 \leq S_{min}$, the optimal w_j will be found on the upper boundary, w_j^u. For values of $S_{min} < S_0 < -\lambda_j$ and $\lambda_j < S_0 < S_{max}$ the optimal w_j is found at the point where the marginal cost equals S_0 whereas no transaction takes place if $|S_0| \leq \lambda_j$.

Therefore, the optimal vector $\hat{\mathbf{w}}$ which minimises the objective:

$$H(\mathbf{w}) = \frac{\kappa}{2} \mathbf{w}^T \mathbf{S} \mathbf{w} - \mathbf{w}^T \bar{\mathbf{x}} + f(\mathbf{d}), \tag{5.4}$$

where

$$f(\mathbf{d}) = \boldsymbol{\lambda}^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathbf{K}\mathbf{d}$$

subject to the (co-ordinatewise) boundary constraints $\mathbf{w}^l \leq \mathbf{w} \leq \mathbf{w}^u$ and the full investment constraint $\mathbf{1}^T\mathbf{w} = 1$ is given by the following algorithm:

Inequality Constrained Shooting Algorithm

1. Start with a value ϕ_0 for the Lagrange multiplier.

2. Start the iteration with an estimator $\hat{\mathbf{w}}_0$

3. For each $j = 1, \ldots, N$ set

$$S_{max} = \lambda_j + (m_j + k_j)(w_{j,t-1} - w_j^l)$$

$$S_{min} = -\lambda_j - (m_j + k_j)(w_j^u - w_{j,t-1})$$

4. At step i, let

$$S_0 = S_j(w_{j,t-1}, \hat{\mathbf{w}}_{i-1}^{-j}, \phi_{k-1})$$

   and set

$$\hat{w}_j = \begin{cases} w_j^l & \text{if } S_0 \geq S_{max} \\ w_{j,t-1} + \frac{\lambda_j - S_0}{m_j + k_j} & \text{if } \lambda_j < S_0 < S_{max} \\ w_{j,t-1} & \text{if } |S_0| \leq \lambda_j \\ w_{j,t-1} + \frac{-\lambda_j - S_0}{m_j + k_j} & \text{if } S_{min} < S_0 < -\lambda_j \\ w_j^u & \text{if } S_0 \leq S_{min} \end{cases}$$

   Form a new estimator $\hat{\mathbf{w}}_i = (\hat{w}_1, \ldots, \hat{w}_N)^T$ after all $\hat{w}_j$'s have been updated.

5. Repeat Step 4 until $\hat{\mathbf{w}}_i$ converges.

6. Update ϕ by using $\phi_k = \phi_{k-1} + 2^{1-k}\,\alpha\,\mathrm{sgn}\,(\mathbf{1}^T\hat{\mathbf{w}}_i - 1)$

7. Repeat Steps 4-6 until convergence of ϕ_k.

Obviously, the modified shooting algorithm can be recovered by setting the k_j's to zero, and removing the portfolio constraints. On the other hand, setting both λ_j and k_j to zero, we obtain an alternative optimization algorithm for a linear programming problem.

5.5 Using the Quadratic Utility Function

We investigate in this section a deviation from the negative exponential utility family and demonstrate how this method would work under an alternative utility function. As mentioned in the first chapter, using the general class of utility functions in equation (1.4) with $b = 1$ and $\gamma = 2$ leads to the quadratic utility function for a portfolio return R_t which can be further simplified to:

$$Q(R_t) = R_t - \frac{\beta}{2} R_t^2.$$

Using the same approximation as before $R_t \approx \mathbf{w}^T \mathbf{x}_t - \lambda\, \mathbf{1}^T \mathbf{d}$ and taking the expected utility leads to:

$$\begin{aligned} E\left[Q(R_t)\right] &= E[R_t] - \frac{\beta}{2} E[R_t^2] \\ &= \mathbf{w}^T \boldsymbol{\mu} - \lambda\, \mathbf{1}^T \mathbf{d} - \frac{\beta}{2} \left[\left(\mathbf{w}^T \boldsymbol{\mu} - \lambda\, \mathbf{1}^T \mathbf{d} \right)^2 + \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \right] \\ &= \mathbf{w}^T \boldsymbol{\mu} - \lambda\, \mathbf{1}^T \mathbf{d} - \frac{\beta}{2} \left[\mathbf{w}^T (\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \mathbf{w} - \lambda\, (\mathbf{1}^T \mathbf{d})(2\, \mathbf{w}^T \boldsymbol{\mu} - \lambda\, \mathbf{1}^T \mathbf{d}) \right] \\ &= -\frac{\beta}{2} \mathbf{w}^T (\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \mathbf{w} + \eta_1\, (\mathbf{w}^T \boldsymbol{\mu}) - \lambda\, \eta_2\, (\mathbf{1}^T \mathbf{d}), \end{aligned}$$

where $\eta_1 = \left(1 + \lambda\beta\, \mathbf{1}^T \mathbf{d}\right) \geq 1$ and $\eta_2 = \left(1 + \frac{1}{2}\lambda\beta\, \mathbf{1}^T \mathbf{d}\right) \geq 1$. The optimal portfolio vector, $\mathbf{w}$, is chosen by maximising the sample equivalent of the expected utility. This is the same as minimising (including the Lagrange multiplier):

$$H(\mathbf{w}, \phi) = -E[Q(R_t)] + \phi\, (\mathbf{1}^T \mathbf{w} - 1).$$

In practice we will be working with the sample equivalent of $H(\mathbf{w}, \phi)$. Then, for a set value of $d^* = \mathbf{1}^T \mathbf{d}$, differentiation with respect to each w_j for $j = 1, \ldots, N$ results in:

$$\frac{\partial H}{\partial w_j} = m_j w_j + c_j + \lambda\, \eta_1^* \operatorname{sgn}(d_j),$$

where

$$m_j = \beta \left(\frac{\mathbf{x}_j^T \mathbf{x}_j}{T} - \lambda\, \bar{x}_j \,\mathrm{sgn}\,(d_j) \right)$$

and

$$c_j = \phi - \eta_1^* \bar{x}_j + \beta \sum_{i \neq j} \left(\frac{\mathbf{x}_j^T \mathbf{x}_i}{T} - \lambda\, \bar{x}_i \,\mathrm{sgn}\,(d_j) \right) w_i.$$

In the above equations, η_1^* is the value of η_1 at $\mathbf{1}^T\mathbf{d} = d^*$. This parameter directly influences the effect of transaction costs on the optimization.

Unlike the negative exponential utility case, as a result of $\mathbf{d}$ being a function of the final portfolio weights, the optimization process will require an extra step. For its computation we need to consider

$$m_j w_j + c_j + \lambda\, \eta_1^* \,\mathrm{sgn}\,(d_j) = 0, \qquad j = 1, \ldots, N$$

and assuming $m_j > 0, \forall j$ there will be a unique set of solutions. A further difference with the negative exponential utility case is that the gradient m_j takes two values for the two cases $\mathrm{sgn}\,(d_j) = \pm 1$. Typically, $m_j > 0$ for both cases and hence there will be only one solution as Figure 5.3 shows.

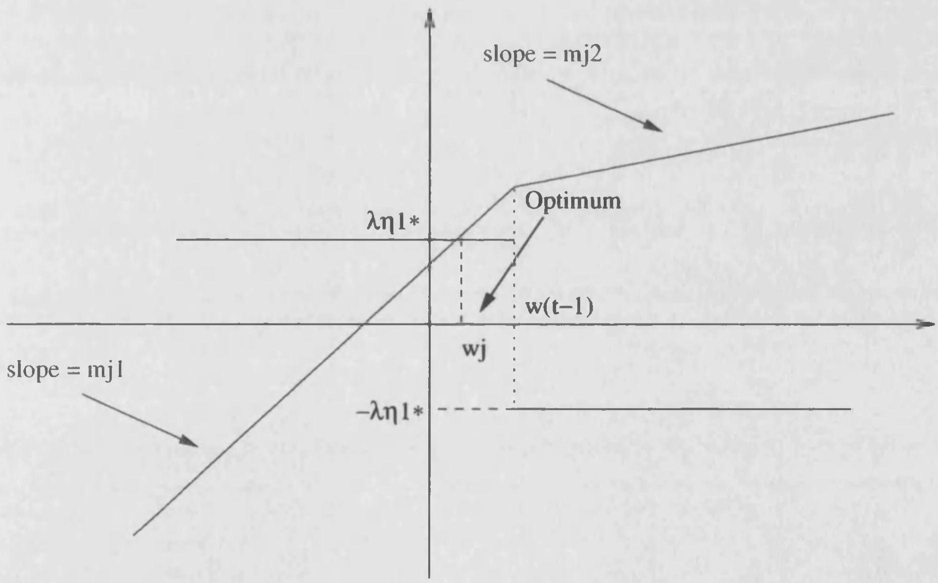

Figure 5.3: The Shooting Method under Quadratic Utility

The iterative process may start with $d^* = 0$ resulting in $\eta_1^* = 1$. Once optimization has been achieved and the Lagrangian has converged so that

$\mathbf{1}^T\mathbf{w} = 1$ (as in the case under the negative exponential utility function), d^* can be updated with the new $\mathbf{w}$. A new d^*, and thus η_1^*, will lead to new optimal weights $\mathbf{w}$. The process can be repeated until $\mathbf{w}$, and hence d and η_1, converge.

It is interesting to highlight the effect of a unit change in d^* in the iterative process. As the total costs increase, there are two reactions. The first concerns the horizontal lines of Figure 5.3. These will be pushed outwards as a direct effect of an absolute increase in $\lambda\,\eta_1^*\,\mathrm{sgn}\,(d_j)$ by a magnitude of $\lambda^2\beta$ for each unit change. At the same time, c_j will change according to the sign of $\bar{x}_j$ by $-\lambda\beta\bar{x}_j$ units. Using as a point of reference the no-cost vector $\mathbf{w}_{t-1}$, the total effect on the optimal w_j will depend on the the magnitude and sign of $\bar{x}_j$. If the sample mean and therefore the expected return is large, the change in c_j will dominate the change in $\lambda\,\eta_1^*\,\mathrm{sgn}\,(d_j)$, thus prompting the investor to increase w_j as it would be worthwhile. On the other hand if $\bar{x}_j$ is negative, the iteration has the opposite result while for small $\bar{x}_j$, the effect of the increase in transaction costs will be the dominating factor.

If the two values for each m_j are of different sign then, there could be 0, 1 or 2 solutions for w_j. If there are no solutions we can take $w_j = w_{j,t-1}$ in that iteration. Otherwise, we take the one closest to $w_{j,t-1}$ between the two solutions.

5.6 A Worked Example

This section illustrates through the use of a worked example the effect of transaction costs on the investment decision. The inequality constrained shooting algorithm was used for all variants of optimization objective (5.4) with $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ replacing $(\bar{\mathbf{x}}, \mathbf{S})$, for different choices of parameters $\left\{\boldsymbol{\lambda}, \mathbf{w}^l, \mathbf{w}^u\right\}$. The matrix $\mathbf{K}$ will be assumed to be a matrix of zeroes throughout, while the starting portfolio is set to the equally weighted portfolio i.e $\mathbf{w}_{t-1} = (0.25, 0.25, 0.25, 0.25)^T$. The annualised mean and variance used for this example are given below:

$$\boldsymbol{\mu} = (0.04, 0.03, 0.02, -0.005)^T$$

and

$$\boldsymbol{\Sigma} = \operatorname{diag}(0.30^2, 0.24^2, 0.20^2, 0.16^2)\,.$$

The value of κ for which the Sharpe ratio and the expected utility optimization result in the same tangency portfolio is

$$\kappa = \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = 1.27$$

hence resulting in optimal tangency portfolio

$$\hat{\mathbf{w}}_1 = (0.35, 0.41, 0.39, -0.15)^T \quad \{\boldsymbol{\lambda} = \mathbf{0}\}\,.$$

The unconstrained tangency portfolio is recovered by assuming that $\boldsymbol{\lambda}$ is a vector of zeros whereas $\mathbf{w}^l$ and $\mathbf{w}^u$ do not exist. Setting $\mathbf{w}^l = \mathbf{0}$ and $\mathbf{w}^u = \mathbf{1}$ and repeating the optimization results in

$$\hat{\mathbf{w}}_2 = (0.32, 0.36, 0.32, 0.00)^T \quad \left\{\boldsymbol{\lambda} = \mathbf{0}, \mathbf{w}^l = \mathbf{0}, \mathbf{w}^u = \mathbf{1}\right\},$$

which is the solution which prohibits short-sales.

We will now include both transaction costs and inequality constraints. In this example, we assume that the transaction cost vector $\boldsymbol{\lambda}$ is constant across all assets, i.e. $\boldsymbol{\lambda} = \lambda\,\mathbf{1}$ with $\lambda = 0.005$. Using the same $\left(\mathbf{w}^l, \mathbf{w}^u\right)$ as before we have:

$$\hat{\mathbf{w}}_3 = (0.31, 0.35, 0.31, 0.02)^T \quad \left\{\boldsymbol{\lambda} = 0.005 \times \mathbf{1}, \mathbf{w}^l = \mathbf{0}, \mathbf{w}^u = \mathbf{1}\right\}.$$

Since the transaction rate is relatively low the effect of the total costs is small and therefore the optimal solution $\hat{\mathbf{w}}_3$ is close to $\hat{\mathbf{w}}_2$, which simply prohibited short-sales. However, setting $\lambda = 0.015$ in $\boldsymbol{\lambda} = \lambda\,\mathbf{1}$ leads to

$$\hat{\mathbf{w}}_4 = (0.25, 0.25, 0.25, 0.25)^T \quad \left\{\boldsymbol{\lambda} = 0.015 \times \mathbf{1}, \mathbf{w}^l = \mathbf{0}, \mathbf{w}^u = \mathbf{1}\right\},$$

and hence any transaction would not be optimal. Portfolios $\hat{\mathbf{w}}_3$ and $\hat{\mathbf{w}}_4$ illustrate how, depending on the starting portfolio, the transaction rate can play a vital role in the investment decisions. Finally, different stocks may be traded at different rates. For example,

$$\{\boldsymbol{\lambda} = (0.015, 0.015, 0.005, 0.005)^T, \mathbf{w}^l = \mathbf{0}, \mathbf{w}^u = \mathbf{1}\}$$

leads to

$$\hat{\mathbf{w}}_4 = (0.25, 0.26, 0.37, 0.12)^T$$

and therefore, the first two assets which have a higher corresponding transaction rate are only marginally traded whereas the investor re-balances the portfolio with respect to the remaining assets.

5.7 Discussion

The purpose of this chapter is two-fold: on the one hand, it attempts to establish a link between a regression estimator for the statistical linear model and a financial decision problem. On the other hand, an iterative algorithm to find the estimator already exists and we modify it in such a way, that it becomes capable of including some necessary conditions and possible extensions to the financial problem.

The lasso estimator (Tibshirani, 1996) is an extension to the ordinary least squares method by imposing some constraints on the sum of the absolute magnitude of the model coefficients. This achieves stability in the out-of-sample performance of the coefficients by reducing their variance at the expense of introducing some bias. The chapter's first objective is met by recognising the fact that the portfolio optimization problem with transaction costs can be written in a similar manner as the least squares objective of the linear model with restrictions. With this viewpoint, the model coefficients are now replaced by the portfolio vector $\mathbf{w}$, whereas the coefficient restrictions are equivalent to the effect of the transaction costs.

In order to achieve the second objective we generalise the shooting algorithm, proposed by Fu (1998), in a number of ways. Firstly we introduce an additional unknown, the Lagrange multiplier, to enforce that the elements of the optimal solution vector $\hat{\mathbf{w}}$ add to 1. This is a necessary constraint for the portfolio selection problem. Secondly, we introduce individual inequality constraints for each element in $\mathbf{w}$. These can be, for example, strictly non-negative thus reflecting the situation faced by an individual investor who is not allowed any short-selling. Alternatively, such constraints may account for possible restrictions faced by an institution or a company with respect to specific stocks. We then proceed to modify the algorithm even further in order to account for the possibility of different proportional transaction costs for each element of $\mathbf{w}$, a situation which could be linked with the liquidity of each stock in the market. Our final contribution is to allow for different proportional costs and increasing marginal (which are not necessarily linear) costs for each asset. Such an extension can be useful to large institutions whose actions in the stock-market may have an effect on the market itself.

Coupled with the capacity of the method working under certain assumptions with different utility functions, these extensions set the framework for a practical and computationally inexpensive procedure of re-balancing a portfolio in the sequential setting. The algorithm can also be applied in other fields where the model coefficients need to satisfy equivalent constraints.

However, it is also important here to note that, in contrast to previous chapters, little attention was paid towards the uncertainty of the optimization inputs, $\bar{\mathbf{x}}$ and $\mathbf{S}$. The reason for this, at first glance, omission is the existence of transaction costs. These impose penalties on the optimization objective, in a similar manner with shrinkage estimators, thus prevent the portfolio from fluctuating widely. The effect of transaction costs is, in a way, analogous to the effect of accounting for uncertainty. Only when the gain of choosing an alternative portfolio is sufficiently large and thus outweighs the transaction costs, will investors re-balance their portfolios. A parallel philosophy was adopted when choosing optimal intensities for the proposed shrinkage estimators. The reduction in in-sample performance was counter-balanced by the reduction in out-of-sample variability.

Chapter 6

A Simulation Study and Empirical Results

In this chapter we will be investigating the behaviour of the proposed estimators under different scenarios and assess their performance against the aforementioned established methods of portfolio selection. Furthermore, historical data from 10 stocks are analysed and portfolio optimization results are presented.

6.1 Estimators' Competition

We start by creating a setup for a competition between different optimal portfolio estimators. The objective is to investigate how the methods proposed by this research would fare against their rivals, under a repeated controlled environment.

6.1.1 Competition Design

A number of different sets of simulation parameters are chosen in such a way to represent possible market conditions. Such conditions include either low or high signal-to-noise ratios or some structure in the true simulation parameters. We restrict ourselves to Normally distributed returns although one could have considered alternative multivariate distributions as the ones found in, for example, Fang et al. (1990) or Bingham and Kiesel (2002).

Once the simulation or *theoretical* parameters are chosen, a number ($L = 1000$) of 'datasets' is generated. The size of each dataset was chosen to be $T = 120$, thus reflecting 10 years of monthly observations. For each simulated dataset, the following 8 estimators are computed:

- Sample tangency portfolio, $\hat{\mathbf{w}}_{SM}$
- Inequality constrained sample tangency portfolio, $\hat{\mathbf{w}}_{QP}$
- James-Stein portfolio, $\hat{\mathbf{w}}_{JS}$
- Ledoit portfolio, $\hat{\mathbf{w}}_{L}$
- Frost-Savarino portfolio, $\hat{\mathbf{w}}_{FS}$
- Britten-Jones portfolio, $\hat{\mathbf{w}}_{BJ}$
- Multivariate L_1 median tangency portfolio, $\tilde{\mathbf{w}}_{SM}$
- Shrinkage $\alpha\beta$ portfolio, $\hat{\mathbf{w}}_{\alpha\beta}$

The calculation of $\tilde{\mathbf{w}}_{SM}$ and $\hat{\mathbf{w}}_{\alpha\beta}$, was based on $B = 100$ bootstrap resamples from each dataset.

Each estimator $\hat{\mathbf{w}}$ results in a theoretical Sharpe Ratio, i.e. the Sharpe Ratio that would have been obtained if the simulation parameters were true. The comparison is based on the distribution of the resulting Sharpe Ratios, for each estimator as well as their relative performance for each simulated dataset.

6.1.2 Scenario 1

Under the first scenario we assume a market governed by uncertainty. The stock volatility is very large relative to the share price movements, resulting in time-series with low 'signal-to-noise' ratios. Such market conditions would be reflected by large elements on the diagonal of the asset return covariance matrix, accompanied by small elements in the expected return vector.

Parameters

For this simulation, the 4 available assets in the stock universe have theoretical excess mean and variance

$$\boldsymbol{\mu} = \frac{1}{12}(0.04, 0.03, 0.02, -0.005)^T$$
$$\boldsymbol{\Sigma} = \frac{1}{12}\,\mathrm{diag}\,(0.30^2, 0.24^2, 0.20^2, 0.16^2)$$

resulting in optimal tangency weights

$$\mathbf{w} = (0.35, 0.41, 0.39, -0.15)^T.$$

The maximum attainable Sharpe ratio is $\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = 0.061$ at the optimal weights. It is interesting to note that this quantity is, in a sense, analogous to the signal-to-noise ratio which is usually encountered in time-series analysis. More specifically, it is a general coefficient of variation for the multivariate asset returns.

Results

The estimators are compared on the basis of the theoretical Sharpe ratio, that they would give rise to. The distribution of each estimator's resulting theoretical Sharpe ratio is plotted in Figure 6.1 below. Alternatively, similar information is conveyed through Table 6.1. For each of the portfolio estimators, a six-figure summary is displayed for the distribution of its theoretical Sharpe ratio. The columns show the minimum, the lower quartile, the mean, the median, the upper quartile and the maximum of the distribution.

A number of conclusions can be drawn from Figure 6.1 and Table 6.1. The sample moments estimator fails to produce portfolios which consistently exhibit high out-of-sample Sharpe ratio. The average theoretical Sharpe ratio is 0.027 whereas within the 1000 simulated datasets, it reached -0.041. The wide range of values reflects the instability of $\hat{\mathbf{w}}_{SM}$.

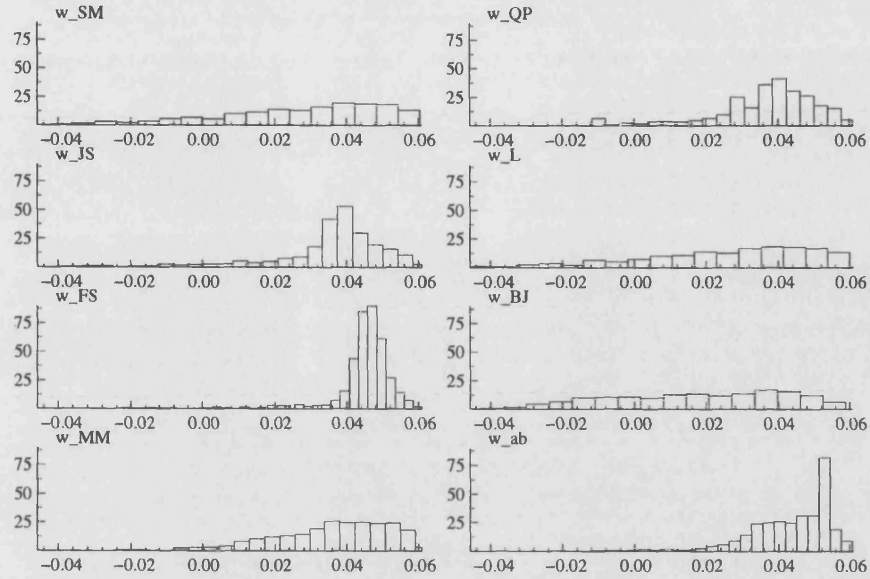

Figure 6.1: Sharpe Ratio Distribution under Scenario 1

Estimator	Min	Q1	Mean	Median	Q2	Max
$\hat{\mathbf{w}}_{SM}$	-0.041	0.013	0.027	0.032	0.046	0.061
$\hat{\mathbf{w}}_{QP}$	-0.009	0.029	0.037	0.038	0.046	0.060
$\hat{\mathbf{w}}_{JS}$	-0.039	0.033	0.035	0.039	0.044	0.060
$\hat{\mathbf{w}}_{L}$	-0.046	0.012	0.026	0.031	0.045	0.061
$\hat{\mathbf{w}}_{FS}$	-0.022	0.043	0.045	0.046	0.049	0.060
$\hat{\mathbf{w}}_{BJ}$	-0.043	-0.001	0.018	0.020	0.038	0.059
$\tilde{\mathbf{w}}_{SM}$	-0.023	0.026	0.036	0.038	0.048	0.061
$\hat{\mathbf{w}}_{\alpha\beta}$	-0.024	0.037	0.043	0.046	0.053	0.060

Table 6.1: Summary Statistics for Sharpe Ratio under Scenario 1

Interestingly, imposing non-negativity constraints ($\hat{\mathbf{w}}_{QP}$) improves the theoretical Sharpe ratio, even though the true tangency portfolio includes an asset with a short position. In fact, the average Sharpe Ratio rises from 0.027 to 0.037 while even higher increases are encountered in the lower quartile and minimum of the distribution. This is also obvious from the respective histograms of the two Sharpe ratio distributions. The optimization constraints limit the uncertainty by forcing the optimization process into selecting "reasonable" portfolios. Furthermore, since the constrained optimization estimator is an extreme case of shrinkage estimation, this result sends early signals about the potential improvement gained by existing and proposed shrinkage estimators.

Current shrinkage estimators ($\hat{\mathbf{w}}_{JS}$, $\hat{\mathbf{w}}_{L}$, $\hat{\mathbf{w}}_{FS}$, $\hat{\mathbf{w}}_{BJ}$) yield mixed results under this scenario. Overall the Frost-Savarino and James-Stein estimators perform better than, whereas the Ledoit estimator similar to, the plug-in sample moments estimator. Somewhat surprisingly, the Britten-Jones estimator seems to perform even worse than $\hat{\mathbf{w}}_{SM}$.

On average the Frost-Savarino estimator results in the highest Sharpe ratio across all estimators (0.045). Moreover, the lower quartile of the distribution for the Sharpe ratio of $\hat{\mathbf{w}}_{FS}$ (0.043) indicates that the Frost-Savarino estimator produces consistently good results.

We now turn our attention to the resulting Sharpe ratio distributions of our proposed estimators ($\tilde{\mathbf{w}}_{SM}$, $\hat{\mathbf{w}}_{\alpha\beta}$). Compared to the sample moments estimator, the average theoretical Sharpe ratio for either proposed estimator is higher and in the case of $\hat{\mathbf{w}}_{\alpha\beta}$ second only to the one achieved by $\hat{\mathbf{w}}_{FS}$. Moreover, both produce consistently high Sharpe ratios with the upper quartile of the Sharpe ratio distribution of $\hat{\mathbf{w}}_{\alpha\beta}$ being the higher across all other estimators.

In order to assess the relative performance of the two proposed estimators we present a head-to-head comparison between all pairs of estimators in Table 6.2. It shows the number of times (out of the total number of simulations) that the row estimator produced a higher theoretical Sharpe ratio than then column estimator. Hence, for example, in 663 simulated datasets, $\hat{\mathbf{w}}_{QP}$ was better than $\hat{\mathbf{w}}_{SM}$ whereas the contrary was true 251 times. Evidently in $1000 - 663 - 251 = 86$ times, the two estimators produced identical results (due to the fact that $\hat{\mathbf{w}}_{SM}$ did not include any short positions and hence was the same as the inequality constrained estimator).

Table 6.2 confirms the conclusions drawn earlier from the summary statis-

	$\hat{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{QP}$	$\hat{\mathbf{w}}_{JS}$	$\hat{\mathbf{w}}_{L}$	$\hat{\mathbf{w}}_{FS}$	$\hat{\mathbf{w}}_{BJ}$	$\tilde{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{\alpha\beta}$
$\hat{\mathbf{w}}_{SM}$	*	251	262	919	206	558	233	113
$\hat{\mathbf{w}}_{QP}$	663	*	514	805	288	751	521	234
$\hat{\mathbf{w}}_{JS}$	738	486	*	753	139	702	472	225
$\hat{\mathbf{w}}_{L}$	81	195	247	*	202	547	186	104
$\hat{\mathbf{w}}_{FS}$	794	712	861	798	*	924	716	496
$\hat{\mathbf{w}}_{BJ}$	442	249	298	453	76	*	300	146
$\tilde{\mathbf{w}}_{SM}$	767	479	528	814	284	700	*	225
$\hat{\mathbf{w}}_{\alpha\beta}$	887	766	760	896	504	854	775	*

Table 6.2: Pairwise comparisons under Scenario 1

tics. For example, the sample moments estimator can be improved upon with almost any shrinkage estimation method. Furthermore, under a market scenario of significant market uncertainty, the Frost-Savarino estimator outperforms almost all other competitors.

Important conclusions regarding the performance of $\tilde{\mathbf{w}}_{SM}$ and $\hat{\mathbf{w}}_{\alpha\beta}$ can also be reached from this pairwise comparison. The shrinkage $\hat{\mathbf{w}}_{\alpha\beta}$ is significantly better than any other estimator and comparable to the Frost-Savarino estimation method. As for the L_1 median estimator $\tilde{\mathbf{w}}_{SM}$, the pairwise comparison with $\hat{\mathbf{w}}_{SM}$ allows us to target the improvement achieved by bootstrap aggregating. More specifically under this market scenario, more than 3 out of 4 times bagging improves the performance of the sample moments estimator. The remaining comparisons for $\tilde{\mathbf{w}}_{SM}$ may be somewhat misleading since a fairer comparison with other estimators would be to calculate their bootstrap aggregated equivalent and investigate the results.

6.1.3 Scenario 2

Under the second scenario we assume a market with lower uncertainty regarding share prices, compared to the first scenario. Alternatively, one can view this scenario as one which would give rise to time-series with high 'signal-to-noise' ratio. This would reflect a situation with investors with a high degree of information about the possible movements of the asset returns.

Parameters

The chosen mean vector and variance-covariance matrix for the second simulation are:

$$\boldsymbol{\mu} = \frac{1}{12}(0.08, 0.06, 0.04, -0.01)^T$$
$$\boldsymbol{\Sigma} = \frac{1}{12}\,\mathrm{diag}\,(0.24^2, 0.18^2, 0.16^2, 0.15^2)$$

resulting in optimal tangency weights

$$\mathbf{w} = (0.32, 0.42, 0.36, -0.10)^T.$$

Under these parameters, the optimal tangency portfolio has a Sharpe ratio of $\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = 0.155$.

Results

The graphical and tabular tools used to compare the estimators are the same as before. The histograms of the distribution of the theoretical Sharpe ratio are shown in Figure 6.2. The six-figure summary for each estimator is presented in Table 6.3.

As before, both the histograms and the table indicate that the sample moments estimator exhibits instability with its out-of-sample performance. In fact, the range of theoretical Sharpe ratio values has increased relative to the previous scenario. This illustrates our assertion that the higher the expected Sharpe ratio of $\hat{\mathbf{w}}_{SM}$ is, the more unstable the estimator becomes.

When the market does not exhibit a lot of uncertainty, as this scenario depicts, shrinkage does not improve the sample moments estimator as dramatically. As before, the Frost-Savarino estimator produces on average the best results among the current methods. Nevertheless, the pairwise comparison between $\hat{\mathbf{w}}_{SM}$ and the remaining current shrinkage methods does not consistently favour one over the rest, which is also evident in Table 6.4.

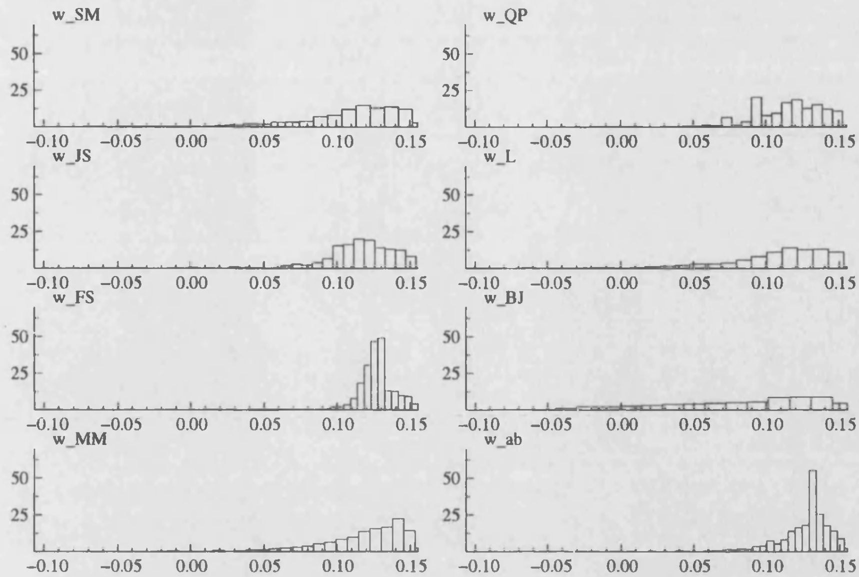

Figure 6.2: Sharpe Ratio Distribution under Scenario 2

Estimator	Min	Q1	Mean	Median	Q2	Max
$\hat{\mathbf{w}}_{SM}$	-0.065	0.093	0.108	0.117	0.135	0.155
$\hat{\mathbf{w}}_{QP}$	-0.019	0.096	0.115	0.118	0.134	0.154
$\hat{\mathbf{w}}_{JS}$	-0.064	0.104	0.115	0.118	0.132	0.155
$\hat{\mathbf{w}}_{L}$	-0.107	0.090	0.106	0.115	0.134	0.155
$\hat{\mathbf{w}}_{FS}$	-0.041	0.121	0.125	0.127	0.132	0.154
$\hat{\mathbf{w}}_{BJ}$	-0.090	0.041	0.080	0.094	0.125	0.155
$\tilde{\mathbf{w}}_{SM}$	-0.020	0.105	0.119	0.125	0.140	0.155
$\hat{\mathbf{w}}_{\alpha\beta}$	0.006	0.120	0.127	0.132	0.137	0.155

Table 6.3: Summary Statistics for Sharpe Ratio under Scenario 2

However, the same cannot be said for the two estimators proposed in this research. Both $\tilde{\mathbf{w}}_{SM}$ and $\hat{\mathbf{w}}_{\alpha\beta}$ result in portfolios with high out-of-sample Sharpe Ratio. For example the bootstrap aggregating of $\hat{\mathbf{w}}_{SM}$ (i.e. $\tilde{\mathbf{w}}_{SM}$) leads to an estimator which outperforms the original sample moments estimator in more than 70% of the simulated samples. Moreover, the proposed shrinkage estimator $\hat{\mathbf{w}}_{\alpha\beta}$ has the highest average theoretical Sharpe ratio

	$\hat{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{QP}$	$\hat{\mathbf{w}}_{JS}$	$\hat{\mathbf{w}}_{L}$	$\hat{\mathbf{w}}_{FS}$	$\hat{\mathbf{w}}_{BJ}$	$\tilde{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{\alpha\beta}$
$\hat{\mathbf{w}}_{SM}$	*	416	302	928	303	581	279	222
$\hat{\mathbf{w}}_{QP}$	451	*	472	643	328	659	402	264
$\hat{\mathbf{w}}_{JS}$	698	528	*	718	257	675	455	232
$\hat{\mathbf{w}}_{L}$	72	357	282	*	287	568	204	197
$\hat{\mathbf{w}}_{FS}$	697	672	743	713	*	873	568	446
$\hat{\mathbf{w}}_{BJ}$	419	341	325	432	127	*	314	211
$\tilde{\mathbf{w}}_{SM}$	721	598	545	796	432	686	*	363
$\hat{\mathbf{w}}_{\alpha\beta}$	778	736	759	803	554	789	637	*

Table 6.4: Pairwise comparisons under Scenario 2

compared to all other competitors. In fact, it is also the only method which led to a positive theoretical Sharpe ratio for all simulated samples. The superior performance of the $\alpha\beta$ estimator is also highlighted by all pairwise comparisons with each of the remaining estimators.

6.1.4 Scenario 3

We consider now an alternative situation which explores the behaviour of the different estimators when the stocks have certain characteristics. More specifically, under the third scenario, we investigate the performance of the estimators when the 4 assets in the universe share the same expected return.

Parameters

Since all shares are chosen to have the same theoretical expected return, $\boldsymbol{\mu}$ will be proportional to a vector of 1s, $\mathbf{1}$. More specifically,

$$\boldsymbol{\mu} = \frac{1}{12}(0.05, 0.05, 0.05, 0.05)^T$$

while the variance-covariance matrix is

$$\boldsymbol{\Sigma} = \frac{1}{12}\,\mathrm{diag}\,(0.24^2, 0.18^2, 0.16^2, 0.15^2)$$

as in the previous scenario. The optimal tangency portfolio will coincide in this case with the global minimum variance portfolio since $\boldsymbol{\mu} \propto \mathbf{1}$. As a result,

$$\mathbf{w} = (0.13, 0.23, 0.30, 0.34)^T.$$

The maximum attainable Sharpe ratio is $\sqrt{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} = 0.166$. Note that this is higher than the previous scenario, since now the assets with the smaller variance exhibit increased expected return (relative to Scenario 2), and the ones with high variance have reduced means.

Results

The histograms of the distribution of out-of-sample Sharpe ratios, shown in Figure 6.3, and the summary statistics Table 6.5 point towards similar

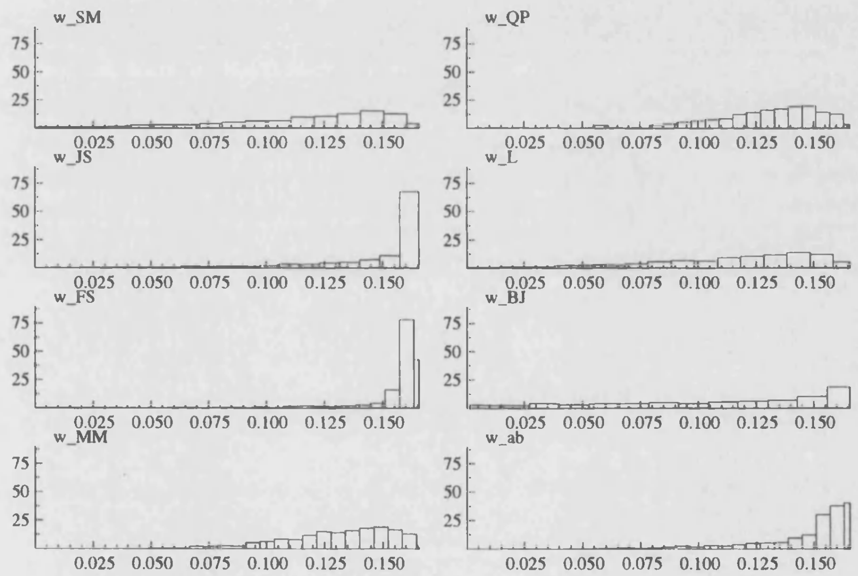

Figure 6.3: Sharpe Ratio Distribution under Scenario 3

conclusions with previous scenarios. Regarding the proposed estimators by this research, the multivariate median of the bootstrap distribution of $\hat{\mathbf{w}}_{SM}$ outperforms the original estimator upon which is based and secondly the $\hat{\mathbf{w}}_{\alpha\beta}$ performs almost as well as the Frost-Savarino method which produces on average the best results.

Estimator	Min	Q1	Mean	Median	Q2	Max
$\hat{\mathbf{w}}_{SM}$	0.001	0.094	0.116	0.125	0.145	0.165
$\hat{\mathbf{w}}_{QP}$	0.060	0.117	0.130	0.134	0.147	0.165
$\hat{\mathbf{w}}_{JS}$	0.001	0.138	0.145	0.160	0.164	0.166
$\hat{\mathbf{w}}_{L}$	0.001	0.090	0.113	0.122	0.143	0.165
$\hat{\mathbf{w}}_{FS}$	0.000	0.158	0.154	0.162	0.164	0.166
$\hat{\mathbf{w}}_{BJ}$	0.001	0.073	0.111	0.124	0.155	0.166
$\tilde{\mathbf{w}}_{SM}$	0.048	0.118	0.131	0.135	0.149	0.165
$\hat{\mathbf{w}}_{\alpha\beta}$	0.020	0.143	0.148	0.156	0.163	0.166

Table 6.5: Summary Statistics for Sharpe Ratio under Scenario 3

	$\hat{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{QP}$	$\hat{\mathbf{w}}_{JS}$	$\hat{\mathbf{w}}_{L}$	$\hat{\mathbf{w}}_{FS}$	$\hat{\mathbf{w}}_{BJ}$	$\tilde{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{\alpha\beta}$
$\hat{\mathbf{w}}_{SM}$	*	116	25	978	48	471	241	20
$\hat{\mathbf{w}}_{QP}$	682	*	172	989	90	561	520	108
$\hat{\mathbf{w}}_{JS}$	975	828	*	974	410	614	893	596
$\hat{\mathbf{w}}_{L}$	22	11	26	*	40	459	114	15
$\hat{\mathbf{w}}_{FS}$	952	910	590	960	*	762	926	671
$\hat{\mathbf{w}}_{BJ}$	529	439	386	541	238	*	466	342
$\tilde{\mathbf{w}}_{SM}$	759	480	107	886	74	534	*	105
$\hat{\mathbf{w}}_{\alpha\beta}$	980	892	393	985	329	658	895	*

Table 6.6: Pairwise comparisons under Scenario 3

An additional interesting conclusion from Table 6.5 and Table 6.6 is related to the nature of the simulation parameters. Recall that the expected return vector $\boldsymbol{\mu}$ for this scenario was chosen to be proportional to a vector of 1s. This possibly explains the performance of the James-Stein method which estimates the mean expected return vector by applying shrinkage of each element of the sample mean towards a grand mean, constant across all assets. Since the shrinkage target coincides with the true mean, methods applying this kind of shrinkage are more likely to result in high out-of-sample Sharpe ratios.

6.1.5 Scenario 4

The fourth scenario that we will be considering enforces some structure on the asset returns' covariance matrix. We now assume that all assets share the same variance and that the covariance matrix is proportional to the identity matrix.

Parameters

The chosen parameters for this simulation study were:

$$\boldsymbol{\mu} = \frac{1}{12}(0.05, 0.04, 0.02, -0.01)^T$$

and

$$\boldsymbol{\Sigma} = \frac{1}{12}\,\mathrm{diag}\,(0.15^2, 0.15^2, 0.15^2, 0.15^2).$$

Since $\boldsymbol{\Sigma} \propto \mathbf{I}$, the optimal tangency weights will be proportional to the expected return vector, $\boldsymbol{\mu}$, and must sum to 1. Hence,

$$\mathbf{w} = (0.5, 0.4, 0.2, -0.1)^T,$$

resulting in maximum Sharpe ratio $\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = 0.131$.

Results

Once again, the histograms of the theoretical Sharpe ratio for each estimation method (Figure 6.4) provide valuable insight in the behaviour of these estimators under this scenario. Similar conclusions may be derived from the summary statistics Table 6.7.

Under the fourth scenario, both the bootstrap aggregating estimator $\tilde{\mathbf{w}}_{SM}$ and the shrinkage estimator $\hat{\mathbf{w}}_{\alpha\beta}$ perform, on average, better than all other estimators. Furthermore, they result in high out-of-sample Sharpe ratio, consistently, since their respective lower quartile of their distributions are at 0.084 and 0.095. From the remaining methods, the Frost-Savarino and

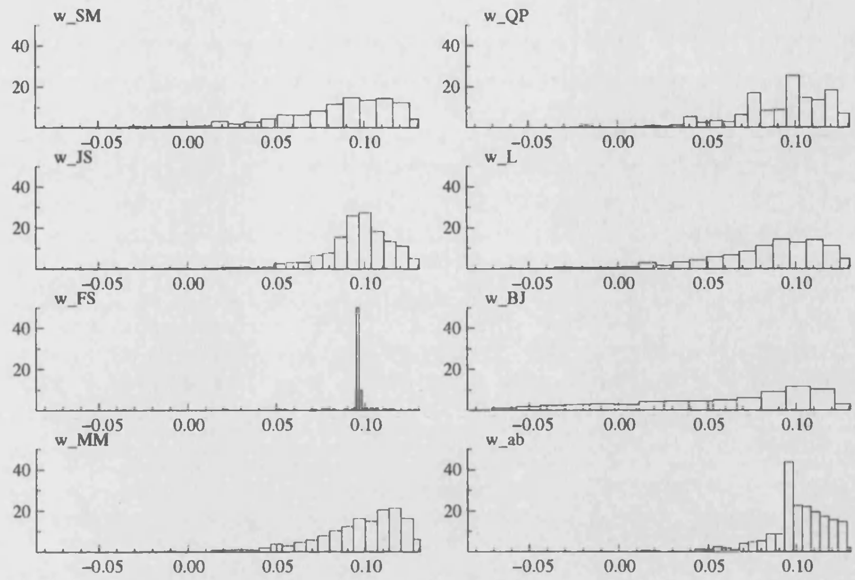

Figure 6.4: Sharpe Ratio Distribution under Scenario 4

Estimator	Min	Q1	Mean	Median	Q2	Max
$\hat{\mathbf{w}}_{SM}$	-0.080	0.067	0.084	0.091	0.110	0.130
$\hat{\mathbf{w}}_{QP}$	-0.019	0.077	0.092	0.096	0.112	0.129
$\hat{\mathbf{w}}_{JS}$	-0.080	0.086	0.094	0.098	0.108	0.130
$\hat{\mathbf{w}}_{L}$	-0.080	0.066	0.082	0.090	0.109	0.130
$\hat{\mathbf{w}}_{FS}$	-0.035	0.096	0.096	0.096	0.096	0.128
$\hat{\mathbf{w}}_{BJ}$	-0.085	0.024	0.060	0.076	0.102	0.130
$\tilde{\mathbf{w}}_{SM}$	-0.023	0.084	0.097	0.102	0.116	0.130
$\hat{\mathbf{w}}_{\alpha\beta}$	-0.029	0.095	0.101	0.101	0.113	0.130

Table 6.7: Summary Statistics for Sharpe Ratio under Scenario 4

James-Stein estimators produce marginally inferior to the aforementioned two whereas other estimators have significantly lower out-of-sample Sharpe ratio.

Interestingly, the distribution of the Frost-Savarino estimator exhibits a peak around one value. Closer investigation revealed that the optimal shrinkage intensity for the sample mean and the variance-covariance matrix for most simulated samples forced the portfolio very close to the equal weighted portfolio, which therefore produced very similar theoretical Sharpe ratios each time.

The pairwise comparison between the estimators in Table 6.8 again confirms the superiority of both $\tilde{\mathbf{w}}_{SM}$ and $\hat{\mathbf{w}}_{\alpha\beta}$ against almost all other estimators. Only the James-Stein method produces comparable results with $\tilde{\mathbf{w}}_{SM}$.

	$\hat{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{QP}$	$\hat{\mathbf{w}}_{JS}$	$\hat{\mathbf{w}}_{L}$	$\hat{\mathbf{w}}_{FS}$	$\hat{\mathbf{w}}_{BJ}$	$\tilde{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{\alpha\beta}$
$\hat{\mathbf{w}}_{SM}$	*	337	249	894	425	544	245	233
$\hat{\mathbf{w}}_{QP}$	545	*	451	679	491	646	407	331
$\hat{\mathbf{w}}_{JS}$	751	549	*	758	542	645	497	314
$\hat{\mathbf{w}}_{L}$	106	321	242	*	407	525	191	217
$\hat{\mathbf{w}}_{FS}$	575	509	458	593	*	680	403	316
$\hat{\mathbf{w}}_{BJ}$	456	354	355	475	320	*	320	217
$\tilde{\mathbf{w}}_{SM}$	755	593	503	809	597	680	*	371
$\hat{\mathbf{w}}_{\alpha\beta}$	767	669	677	783	684	783	629	*

Table 6.8: Pairwise comparisons under Scenario 4

Finally, as in all previous scenarios, the shrinkage estimator $\hat{\mathbf{w}}_{\alpha\beta}$ is better than the bagging estimator $\tilde{\mathbf{w}}_{SM}$.

6.1.6 Scenario 5

We now turn our attention to two scenarios with an increased number of assets. So far all previous simulations dealt with a small number of assets, only suitable for a private investor. A financial institution such as a bank or an individual such as a fund manager will have a portfolio with many more assets. This setting may provide fertile ground for certain estimators which are designed to perform better when a higher number of assets is available. Therefore, we investigate their behaviour and compare it with our proposed methods.

Parameters

We simulated $T = 60$ "monthly" excess returns for each of $N = 30$ assets. Each of the elements in the expected asset returns vector, $\boldsymbol{\mu}$, was a uniform random number in the range $\frac{1}{12}[-0.02, 0.04]$ with its corresponding diagonal element in the simulation covariance matrix $\boldsymbol{\Sigma}$ following a $\mathcal{U}\left(\frac{1}{12}0.18^2, \frac{1}{12}0.30^2\right)$ distribution. All off-diagonal elements in $\boldsymbol{\Sigma}$ were set to 0.

The simulated parameters resulted in a maximum attainable Sharpe Ratio of $\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = 0.149$. In the optimal tangency portfolio, the largest asset allocation was 0.18 whereas the smallest element of $\mathbf{w}$ was -0.05. In total there were 17 positive weights and 13 short positions.

Results

The Sharpe ratio distributions for each of the methods, shown in Figure 6.5 appears more symmetric than in previous simulations. This is also confirmed

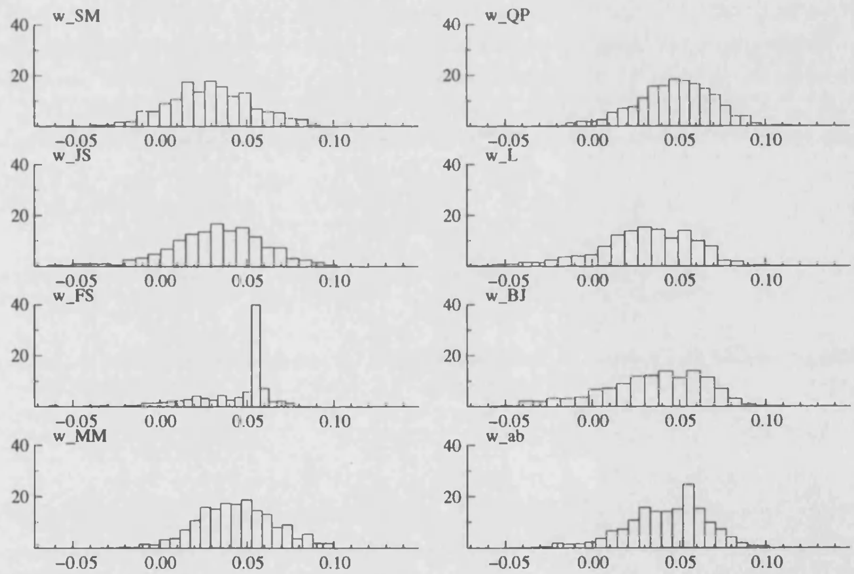

Figure 6.5: Sharpe Ratio Distribution under Scenario 5

by the summary statistics in Table 6.9 showing the mean and the median of the realised Sharpe ratio closer than at previous simulations across most of

Estimator	Min	Q1	Mean	Median	Q2	Max
$\hat{\mathbf{w}}_{SM}$	-0.063	0.014	0.030	0.029	0.046	0.098
$\hat{\mathbf{w}}_{QP}$	-0.022	0.033	0.047	0.049	0.063	0.110
$\hat{\mathbf{w}}_{JS}$	-0.062	0.017	0.033	0.034	0.051	0.102
$\hat{\mathbf{w}}_{L}$	-0.072	0.014	0.030	0.032	0.051	0.103
$\hat{\mathbf{w}}_{FS}$	-0.062	0.057	0.050	0.057	0.057	0.084
$\hat{\mathbf{w}}_{BJ}$	-0.069	0.015	0.034	0.038	0.058	0.103
$\tilde{\mathbf{w}}_{SM}$	-0.034	0.029	0.044	0.043	0.059	0.099
$\hat{\mathbf{w}}_{\alpha\beta}$	-0.064	0.029	0.043	0.045	0.058	0.109

Table 6.9: Summary Statistics for Sharpe Ratio under Scenario 5

the estimators. Moreover, it is interesting to note that none of the methods achieves a relatively high Sharpe ratio. Recall that the optimal tangency portfolio would lead to a Sharpe ratio of 0.149 whereas the inequality constrained estimator merely reaches 0.110.

Based on the summary statistics, on average, the Frost-Savarino estimator results in the highest true Sharpe Ratio, closely followed by the inequality constrained estimator, $\hat{\mathbf{w}}_{QP}$. The proposed estimators $\tilde{\mathbf{w}}_{SM}$ and $\hat{\mathbf{w}}_{\alpha\beta}$ have the next best average Sharpe ratio.

The increased number of assets in the simulation seems to have also resulted in an improved performance for the Ledoit and Britten-Jones estimators, especially relative to the plug-in sample moments estimator. The pairwise comparisons are summarised in Table 6.10 and highlight the fact that these methods are more suited when the number of assets is large. Furthermore, the pairwise comparisons reflect the superiority of the Frost-Savarino estimator and the unexpectedly good results obtained by imposing non-negativity constraints in the optimization. The reason behind the success of $\hat{\mathbf{w}}_{QP}$ lies in the fact that the optimal weights in the tangency portfolio have relatively small magnitude. By forcing the chosen portfolio to consist of only positive elements, the process essentially targets the area of the true optimal

	$\hat{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{QP}$	$\hat{\mathbf{w}}_{JS}$	$\hat{\mathbf{w}}_{L}$	$\hat{\mathbf{w}}_{FS}$	$\hat{\mathbf{w}}_{BJ}$	$\tilde{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{\alpha\beta}$
$\hat{\mathbf{w}}_{SM}$	*	178	183	409	193	331	154	177
$\hat{\mathbf{w}}_{QP}$	822	*	759	796	459	619	589	620
$\hat{\mathbf{w}}_{JS}$	817	241	*	526	285	380	301	378
$\hat{\mathbf{w}}_{L}$	591	204	474	*	266	418	332	302
$\hat{\mathbf{w}}_{FS}$	807	541	715	734	*	648	613	589
$\hat{\mathbf{w}}_{BJ}$	669	381	620	582	352	*	443	477
$\tilde{\mathbf{w}}_{SM}$	846	411	699	668	387	557	*	498
$\hat{\mathbf{w}}_{\alpha\beta}$	817	380	622	698	411	523	502	*

Table 6.10: Pairwise comparisons under Scenario 5

tangency portfolio. As far as the two proposed estimators are concerned, they outperform all remaining methods apart from the aforementioned $\hat{\mathbf{w}}_{FS}$ and $\hat{\mathbf{w}}_{QP}$.

6.1.7 Scenario 6

The sixth and final simulation uses an increased number of available assets ($N = 45$) in order to investigate the behaviour of the eight estimators as the dimension of the problem increases.

Parameters

As before, the number of simulated "datapoints" are twice as many as the assets ($T = 90$), thus simulating $7\frac{1}{2}$ years of monthly returns. The elements of the expected return vector $\boldsymbol{\mu}$ and the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ are randomly chosen from

$$\mu_j \sim \mathcal{U}\left(-\frac{1}{12}0.02, \frac{1}{12}0.04\right), \sigma_j^2 \sim \mathcal{U}\left(\frac{1}{12}0.18^2, \frac{1}{12}0.30^2\right)$$

for $j = 1, \ldots, N$, as in the previous scenario. Once again, the off-diagonal elements of $\boldsymbol{\Sigma}$ are set to 0.

Based on the simulation parameters, the optimal tangency portfolio consists of 18 negative and 27 positive elements, ranging from -0.05 to 0.11, and results in a Sharpe ratio of $\sqrt{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} = 0.162$.

Results

The true Sharpe ratio resulting from each estimator is depicted in Figure 6.6. It is interesting to note that some of the distributions are almost symmetric. The same conclusion can also be reached from the summary statistics in Table 6.11. An additional point of interest regards the maximum attainable

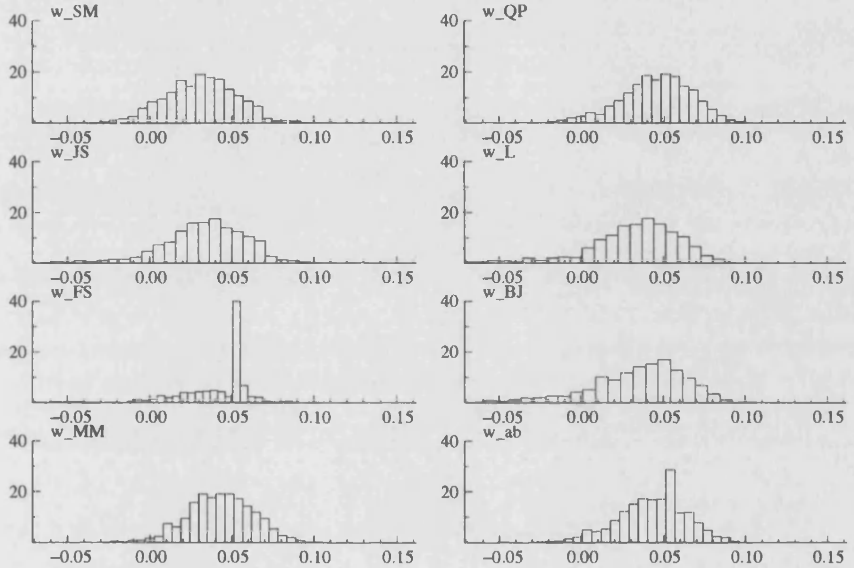

Figure 6.6: Sharpe Ratio Distribution under Scenario 6

Sharpe ratio. None of the estimators achieves a Sharpe ratio which is close to the optimum one, a situation which was also observed in the previous scenario.

The results suggest that, on average, the Frost-Savarino estimator continues to have the best out-of-sample performance. It is also important to note that this estimator produces consistently high results since the lower quartile of its distribution is larger than the median Sharpe ratio of other estimators (and in some cases even higher than their respective upper quartiles).

Estimator	Min	Q1	Mean	Median	Q2	Max
$\hat{\mathbf{w}}_{SM}$	-0.065	0.016	0.031	0.032	0.046	0.095
$\hat{\mathbf{w}}_{QP}$	-0.027	0.032	0.046	0.047	0.060	0.100
$\hat{\mathbf{w}}_{JS}$	-0.065	0.019	0.033	0.035	0.050	0.098
$\hat{\mathbf{w}}_{L}$	-0.071	0.019	0.033	0.036	0.052	0.092
$\hat{\mathbf{w}}_{FS}$	-0.064	0.054	0.048	0.054	0.054	0.097
$\hat{\mathbf{w}}_{BJ}$	-0.067	0.019	0.035	0.040	0.055	0.100
$\tilde{\mathbf{w}}_{SM}$	-0.028	0.030	0.043	0.044	0.057	0.106
$\hat{\mathbf{w}}_{\alpha\beta}$	-0.046	0.030	0.043	0.045	0.056	0.096

Table 6.11: Summary Statistics for Sharpe Ratio under Scenario 6

Both Table 6.11 and the pairwise comparison Table 6.12 point towards the superiority of $\hat{\mathbf{w}}_{FS}$ together with the inequality constrained estimator $\hat{\mathbf{w}}_{QP}$. Among the remaining methods, the estimators $\tilde{\mathbf{w}}_{SM}$ and $\hat{\mathbf{w}}_{\alpha\beta}$ produce

	$\hat{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{QP}$	$\hat{\mathbf{w}}_{JS}$	$\hat{\mathbf{w}}_{L}$	$\hat{\mathbf{w}}_{FS}$	$\hat{\mathbf{w}}_{BJ}$	$\tilde{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{\alpha\beta}$
$\hat{\mathbf{w}}_{SM}$	*	184	167	361	171	302	152	152
$\hat{\mathbf{w}}_{QP}$	816	*	742	728	472	593	595	607
$\hat{\mathbf{w}}_{JS}$	833	258	*	470	299	352	300	357
$\hat{\mathbf{w}}_{L}$	639	272	530	*	321	423	364	340
$\hat{\mathbf{w}}_{FS}$	829	528	701	679	*	634	590	580
$\hat{\mathbf{w}}_{BJ}$	698	407	648	577	366	*	483	494
$\tilde{\mathbf{w}}_{SM}$	848	405	700	636	410	517	*	504
$\hat{\mathbf{w}}_{\alpha\beta}$	834	393	643	660	420	506	496	*

Table 6.12: Pairwise comparisons under Scenario 6

the best results. Relative to $\hat{\mathbf{w}}_{QP}$ and $\hat{\mathbf{w}}_{FS}$, the proposed methods were superior in about 40% of the simulated datasets.

With the larger number of assets in the universe, both the Britten-Jones and Ledoit estimators improved against most of their competitors, a fact which may suggest that their true worth may show with very large N.

6.1.8 Conclusions

These simulations highlight a number of points regarding the lack of stability of the sample moments estimator and the usefulness of alternative methods in selecting optimal portfolios. As an overall comparison, we aggregate the six head-to-head comparison tables (Tables 6.2, 6.4, 6.6, 6.8, 6.10 and 6.12) and present the results in Table 6.13 in percentage terms. In doing so, we

	$\hat{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{QP}$	$\hat{\mathbf{w}}_{JS}$	$\hat{\mathbf{w}}_{L}$	$\hat{\mathbf{w}}_{FS}$	$\hat{\mathbf{w}}_{BJ}$	$\tilde{\mathbf{w}}_{SM}$	$\hat{\mathbf{w}}_{\alpha\beta}$
$\hat{\mathbf{w}}_{SM}$	*	24.70	19.80	74.82	22.43	46.45	21.73	15.28
$\hat{\mathbf{w}}_{QP}$	66.32	*	51.83	77.33	35.47	63.82	50.57	36.07
$\hat{\mathbf{w}}_{JS}$	80.20	48.17	*	69.98	32.20	56.13	48.63	35.03
$\hat{\mathbf{w}}_{L}$	25.18	22.67	30.02	*	25.38	49.00	23.18	19.58
$\hat{\mathbf{w}}_{FS}$	77.57	64.53	67.80	74.62	*	75.35	63.60	51.63
$\hat{\mathbf{w}}_{BJ}$	53.55	36.18	43.87	51.00	24.65	*	38.77	31.45
$\tilde{\mathbf{w}}_{SM}$	78.27	49.43	51.37	76.82	36.40	61.23	*	34.43
$\hat{\mathbf{w}}_{\alpha\beta}$	84.38	63.93	64.23	80.42	48.37	68.55	65.57	*

Table 6.13: Overall pairwise comparisons

are implicitly assuming that all six scenarios are equally likely to occur in real-life situations.

Table 6.13 shows the percentage of simulated datasets for which the row estimator outperformed the column estimator. So for example, out of 6000 simulated datasets, in 80.20% (or in 4812 datasets), the James-Stein estimator $\hat{\mathbf{w}}_{JS}$ produced a higher out-of-sample Sharpe ratio than the sample moments estimator $\hat{\mathbf{w}}_{SM}$. The contrary was true 19.80% of the time. As mentioned before, the two percentages may not sum to 100% as in the case of $\hat{\mathbf{w}}_{SM}$ and $\hat{\mathbf{w}}_{QP}$ due to ties.

Across all scenarios, in varying degrees, the instability of the sample moments estimator was evident. As a result most estimators produced higher

Sharpe ratios than $\hat{\mathbf{w}}_{SM}$. Somewhat, surprisingly the Ledoit estimator performed significantly worse. However, judgement should be reserved for this method as it requires in real-life the returns of a single factor model whereas, in these simulations, a simple average of the returns was used. Furthermore, a breakdown of the performance of $\hat{\mathbf{w}}_L$ against the sample moments plug-in estimator by scenario, shows that it performs significantly better when the number of available assets, N, is large.

The performance of the inequality constrained estimator $\hat{\mathbf{w}}_{QP}$ illustrated how $\hat{\mathbf{w}}_{SM}$ can be improved by even imposing the wrong constraints. The true portfolio included short positions in 5 out of 6 scenarios, yet $\hat{\mathbf{w}}_{QP}$ outperformed $\hat{\mathbf{w}}_{SM}$ under all scenarios. In fact, when the two estimators did not produce identical results, imposing portfolio constraints led to a higher out-of-sample Sharpe ratio more than twice more often.

It can also be concluded that, out of the four current shrinkage methods, the Frost-Savarino estimator $\hat{\mathbf{w}}_{FS}$ produces better results relative to all other existing estimators. However, this significantly superior performance is not reproduced when $\hat{\mathbf{w}}_{FS}$ is compared with the $\hat{\mathbf{w}}_{\alpha\beta}$ shrinkage estimator. In fact, our proposed method performs equally well with $\hat{\mathbf{w}}_{FS}$ and significantly better than all other estimators. Equally important is the fact that this pattern is observed across most market scenarios thus suggesting that investors who choose their portfolios according to this selection procedure will usually fare better than investors following anyone of the other portfolio selection methods.

Finally, using the L_1 median ($\tilde{\mathbf{w}}_{SM}$) of the bootstrap distribution of $\hat{\mathbf{w}}_{SM}$ is three times more likely to improve the original estimator. This was also evident in all of the investigated scenarios. The performance of $\tilde{\mathbf{w}}_{SM}$ against the remaining estimators had mixed success and failure. However, the reason for such mixed fortunes lies in the fact that this method improves an existing estimator. Hence if we wanted to improve $\hat{\mathbf{w}}_{FS}$, we could attempt calculating the multivariate L_1 median of its bootstrap distribution and compare the resulting estimator with $\hat{\mathbf{w}}_{FS}$ over many simulated datasets.

6.2 A Case Study

In this section, we will be investigating the behaviour of the methods in this thesis when applied to a real dataset. Towards this end, the dataset will be split in two: a training set and a test set. The former will be used for the calculation of initial estimators. The first observation from the test set will be used to calculate the return of each portfolio under the different estimators, before subsequently being included in the updated training set. This is repeated for all observations in the test set. Since the true parameters of the data-generating process are unknown, we will compare the methods on the basis of their realised Sharpe Ratio.

6.2.1 The Method

Let us assume that following the close prices at time t, the training sample set $\mathbf{x}_1, \ldots, \mathbf{x}_t$ has been observed. These are $N \times 1$ column vectors of excess returns. Based on predictions of the returns' covariance matrix $\hat{\mathbf{\Sigma}}_{t+1}$ and the expected return vector $\hat{\boldsymbol{\mu}}_{t+1}$, each method results in a portfolio vector $\hat{\mathbf{w}}_{t+1}$. Let us further assume that current portfolio holdings are represented by $\mathbf{w}_t$ whereas current wealth is denoted by M_t. In order to transform portfolio $\mathbf{w}_t$ to $\hat{\mathbf{w}}_{t+1}$, the investor must incur transaction costs C_t such that

$$C_t = \lambda\, M_t\, \mathbf{1}^T \mathbf{d}_t$$

where

$$\mathbf{d}_t = \left\{ |\hat{w}_{j,t+1} - w_{j,t}| : j = 1, \ldots, N \right\}.$$

We are implicitly assuming that transactions can be made before any change in prices at $t+1$.

Once $\mathbf{x}_{t+1}$ is observed, the actual holdings will be $M_t\, \hat{\mathbf{w}}_{t+1} \odot (\mathbf{1} + \mathbf{x}_{t+1})$ where $\odot$ denotes the Hadamard (element-by-element) product of two vectors or matrices. Hence, total wealth is given by

$$M_{t+1} = M_t \left(1 + \hat{\mathbf{w}}_{t+1}^T \mathbf{x}_{t+1}\right)$$

whereas the new portfolio holdings will be given by

$$\begin{aligned}\mathbf{w}_{t+1} &= \frac{\hat{\mathbf{w}}_{t+1} \odot (\mathbf{1} + \mathbf{x}_{t+1})}{\mathbf{1}^T \left[\hat{\mathbf{w}}_{t+1} \odot (\mathbf{1} + \mathbf{x}_{t+1})\right]} \\ &= \frac{\hat{\mathbf{w}}_{t+1} \odot (\mathbf{1} + \mathbf{x}_{t+1})}{\left(1 + \hat{\mathbf{w}}_{t+1}^T \mathbf{x}_{t+1}\right)}.\end{aligned}$$

Finally the after cost portfolio return will be

$$R_{t+1} = \frac{M_{t+1} - M_t - C_t}{M_t}.$$

We repeat this procedure until we obtain the series $\{R_i : i = t+1, \ldots, T\}$ for each method and finally, compare the estimators on the basis of the realised Sharpe ratio given by:

$$\mathrm{RSR}\left[\hat{\mathbf{w}}\right] = \frac{\bar{R} - r}{\mathrm{SD}\left[R\right]}$$

where r is the risk-free rate, $\mathrm{SD}\left[R\right]$ is the standard deviation of the portfolio return given by

$$\mathrm{SD}\left[R\right] = \left[\frac{1}{T-t-1} \sum_{i=t+1}^{T} \left(R_i - \bar{R}\right)^2\right]^{\frac{1}{2}}$$

and

$$\bar{R} = \frac{1}{T-t} \sum_{i=t+1}^{T} R_i$$

denotes the average return.

6.2.2 The Data

We have obtained $T = 1000$ daily closing prices of $N = 10$ assets from the London Stock Exchange. These share prices cover a period of 4 years (1/1/98 – 31/12/01) and were freely available from the Internet[1].

Table 6.14 shows the 10 selected stocks[2] which are also included in the FTSE 100 index. They come from one of four economy sectors. Figure 6.7 shows the share price indices of assets from the Transport sector and Banks

[1] Data available from Yahoo! Finance at http://uk.yahoo.finance.com.

[2] For GAA, there were two missing values which were imputed by linear interpolation.

Symbol	Name	Sector
BAA	BAA	Transport
BARC	BARCLAYS	Banks
BAY	BRITISH AIRWAYS	Transport
BSY	B SKY B	Media
GAA	GRANADA	Media
HSBA	HSBC HOLDINGS	Banks
LLOY	LLOYDS TSB GRP	Banks
RBS	ROYAL BANK SCOT	Banks
RR	ROLLS ROYCE	Aerospace and Defence
RTR	REUTERS GROUP	Media

Table 6.14: Available Stocks in the dataset

sector, over the period of interest. The first observation of each series was used as the base index 100 for better illustrative comparisons. Figure 6.8 plots the share price indices for the remaining four assets from the Aerospace and Defence sector and the Media sector.

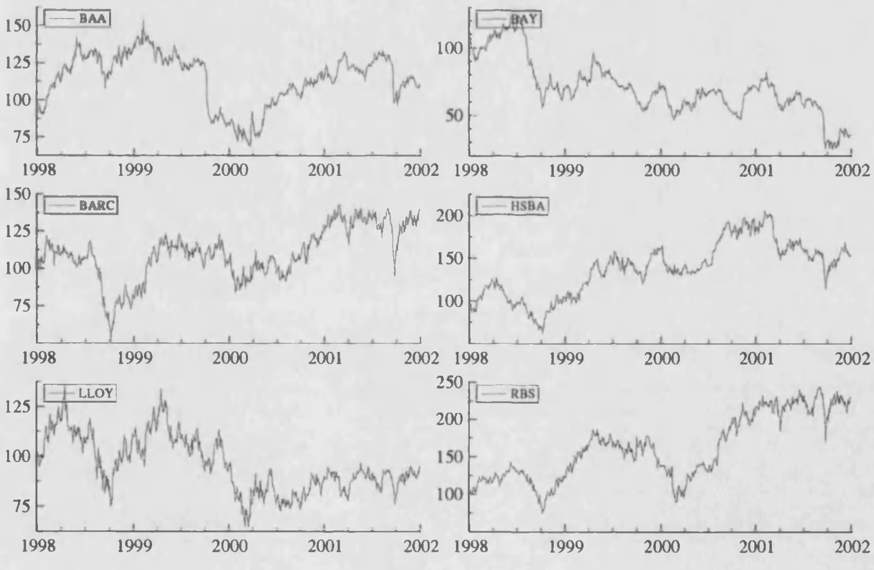

Figure 6.7: Share Price Indices for Transport and Banks

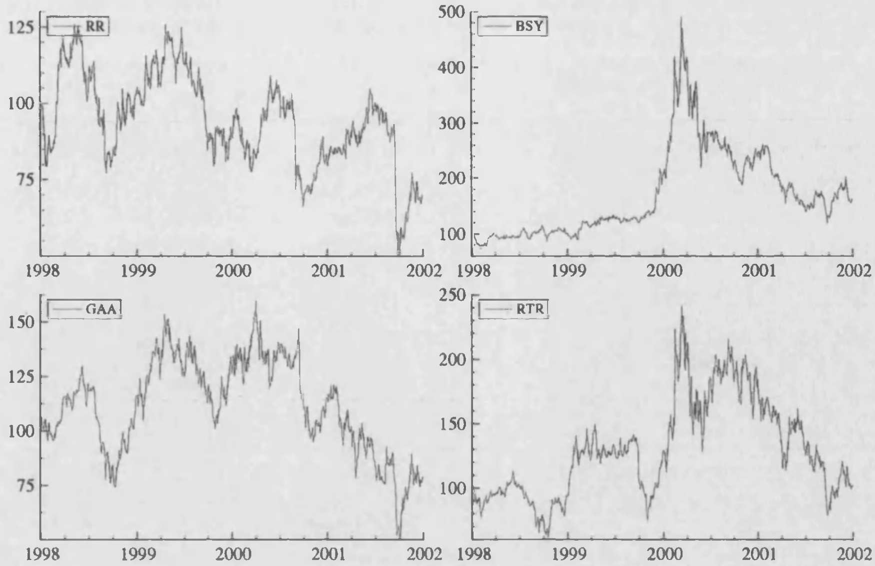

Figure 6.8: Share Price Indices for Aerospace and Defence and Media

Relative returns were obtained and were transformed to excess returns by subtracting the risk-free rate, assumed to be $r = 4\%$ per year. The dataset was split in two groups with the training set consisting of $t = 500$ observations. The remaining datapoints were used in the test set.

Ten estimators were used. Apart from the eight estimators used in simulations earlier in this chapter, the following two were also included in the analysis:

- Lasso estimator with transaction costs, $\hat{\mathbf{w}}_{LS}$
- Generalized shrinkage portfolio, $\hat{\mathbf{w}}_{\lambda_1 \lambda_2}$

The latter was computed with the help of a partly specified transition matrix

$$\mathbf{V} = \begin{pmatrix} 1.0 & & & \\ & 1.0 & & \\ & & 0.1 & \\ & & & 0.01 \end{pmatrix}$$

and $\mathbf{\Sigma}_z = 0.01 \times \mathbf{I}$, based on notation defined in Chapter 4. The vector $\mathbf{w}_0$ was set to be the equally weighted portfolio, i.e $\mathbf{w}_0 = \frac{1}{N}\mathbf{1}$.

6.2.3 Results

Three scenarios are explored. The transaction costs rate was set to $\lambda = 0\%$, 0.5% and 1%. Table 6.15 shows the Realised Sharpe Ratio (RSR) and the final, after-cost wealth, computed with initial wealth $M_t = 1$ for the three scenarios. Even when $\lambda = 0\%$ some estimators $(\hat{\mathbf{w}}_{SM}, \hat{\mathbf{w}}_L, \hat{\mathbf{w}}_{BJ})$ result in

	$\lambda = 0\%$		$\lambda = 0.5\%$		$\lambda = 1\%$	
	RSR	$M_T - \sum C_i$	RSR	$M_T - \sum C_i$	RSR	$M_T - \sum C_i$
$\hat{\mathbf{w}}_{SM}$	*	0	*	0	*	0
$\hat{\mathbf{w}}_{QP}$	0.011	1.016	0.003	0.864	-0.006	0.711
$\hat{\mathbf{w}}_{JS}$	0.006	1.060	-0.006	0.958	-0.017	0.855
$\hat{\mathbf{w}}_L$	*	0	*	0	*	0
$\hat{\mathbf{w}}_{FS}$	0.010	1.101	0.001	1.020	-0.008	0.938
$\hat{\mathbf{w}}_{BJ}$	*	0	*	0	*	0
$\tilde{\mathbf{w}}_{SM}$	-0.016	0.737	*	0	*	0
$\hat{\mathbf{w}}_{\alpha\beta}$	0.021	1.211	*	0	*	0
$\hat{\mathbf{w}}_{\lambda_1\lambda_2}$	0.008	1.056	*	0	*	0
$\hat{\mathbf{w}}_{LS}$	0.011	1.016	0.009	0.974	0.004	0.893

Table 6.15: Performance Measures for Three Scenarios

bankruptcy. Amongst the remaining methods, $\hat{\mathbf{w}}_{\alpha\beta}$ results in the highest realised Sharpe Ratio, and at the same time the highest final, after-cost, wealth. It is followed by the Frost-Savarino method, in terms of wealth and by either $\hat{\mathbf{w}}_{QP}$ or $\hat{\mathbf{w}}_{LS}$ in terms of RSR. It is worth noting that $\hat{\mathbf{w}}_{QP}$ and $\hat{\mathbf{w}}_{LS}$ give identical results since, for $\lambda = 0\%$ both methods solve the same optimization problem in different ways.

When transactions incur costs, with for example $\lambda = 0.5\%$, most methods result in bankruptcy because of the amount of trades. Unfortunately, this is also the fate of both $\hat{\mathbf{w}}_{\alpha\beta}$ and $\hat{\mathbf{w}}_{\lambda_1\lambda_2}$. In terms of realised Sharpe ratio, the lasso estimator outperforms the remaining methods whereas $\hat{\mathbf{w}}_{FS}$ is the only one which results in an overall profit. As the transaction cost rate increases to $\lambda = 1\%$, the situation remains the same. The Frost-Savarino estimator performs marginally better in terms of final wealth, whereas the $\hat{\mathbf{w}}_{LS}$ has the highest RSR.

6.2.4 Conclusions

It is important here to comment on some points that are raised from this analysis. First of all, the sensitivity of some estimators, such as $\hat{\mathbf{w}}_{SM}$, to optimization inputs leads, as has been argued before, to fluctuations in the selected portfolio weights. Extreme, either long or short, positions can be regarded as wild bets which often lead to bankruptcy. This is evident even in the absence of transaction costs. Under this scenario, earlier findings regarding the suitability of our proposed methods are confirmed.

When transactions are penalised with costs proportional to the value of traded stocks (or what is known as *turnover*), more methods lead to bankruptcy. This is the direct consequence of the fact that all estimators bar one are not designed to take into account transaction costs. The lasso estimator is the only one which includes in the optimization objective the effect of transaction costs and, as a result, produces the highest realised Sharpe ratio. Amongst the remaining methods, $\hat{\mathbf{w}}_{FS}$ is the one which leads to the highest final net wealth. It could be argued that even though net wealth may be considered as more appealing, the Sharpe ratio is a measure of return per unit of risk and, therefore, may be regarded as an indicator of consistent future earnings.

An important point that should not be overlooked is the fact that all of the examined estimators are in fact "blind". Optimal portfolios are selected on the basis of historical performance rather than future predictions. Consequently, most fail underlying the fact that better prediction techniques are necessary so that optimization inputs reflect more closely the expected asset returns and covariance matrix.

Overall, this analysis highlights two issues. On the one hand, it is necessary for investors to account for transaction costs when selecting their portfolios. On the other hand, estimation uncertainty does play an important role in portfolio optimization, confirming earlier findings.

Chapter 7

Conclusions

Every day, investors across the world make decisions regarding their portfolio holdings. These decisions are related to investors' expectations on how the market will behave through the population moments of asset returns. Investors' beliefs are usually guesses or estimates of the necessary inputs of the portfolio optimization problem. These estimates which are, at least partly, calculated from historical data involve estimation error which is therefore transferred to the optimal decisions. In some case, the resulting decisions are highly sensitive to the input data, a fact that renders the optimization output as highly unstable in out-of-sample situations. An example of such a case is the sample tangency portfolio which maximizes the expected excess return per unit of risk of a portfolio. Our objective was to develop methodologies which account for the uncertainty governing the estimated inputs in the optimization process and thus result in decisions with better out-of-sample performance.

We achieve a significant improvement in the performance of existing estimators by simply considering what could have been observed in a data sample rather than just what was actually encountered. This is because an estimator may be the optimal, under some criterion, decision to take given the observed sample but in certain cases, there is no guarantee that the same estimator will perform equally well had an alternative dataset been encountered. This is true even if the data-generating process remains the same or even when an alternative sample differs only slightly from the sample on which the estimator was based. Hence, since investors make decisions about the future based on observed samples, we advocate that they should also take into account additional possible samples that could have been obtained by the same data-generating process. We generate such resamples using the bootstrap.

We have found that a robust measure of multivariate location of the bootstrap distribution of the sample moments tangency portfolio is a much improved version of the original sample moments estimator. A similar methodology can be adopted for the improvement of other existing estimators, which may involve constraints on assets holdings, with varying degrees of success.

The improvement achieved with the new estimator is linked with the instability of the original estimator. If the latter is highly unstable in the sense that small changes in the data may lead to dramatic changes in the estimator as in the case of the sample moments tangency portfolio, the improvement will be significant. On the other hand, if the original estimator is well-behaved, the proposed methodology will lead to an estimator with comparable out-of-sample performance.

Existing estimators attempt to solve the problem of instability through shrinkage. The estimated sample moments are pulled towards more stable, desirable, targets with the optimal shrinkage intensity chosen by different methods. Our philosophy remains the same. We propose a way of choosing the shrinkage parameters by accounting for the possible out-of-sample behaviour of the estimator. The chosen optimality criterion rewards a consistently high out-of-sample performance, by achieving a balance between a reduction in variability (and thus instability) and an increase in bias. As before, we generate bootstrap resamples from the original sample to assess the performance of the estimator over a range of the shrinkage intensity parameter values.

Based on the shrinkage approach, we take a further step by modifying the estimator to allow for extensions. The targets towards which shrinkage is applied are now altered. Possible market information or investor preferences may be included under a general framework which allows for these beliefs to have a direct effect on the shrinkage targets. One such example is given to accommodate investors who may prefer to hold approximate proportions of their wealth in pre-specified industries. Once certain tuning coefficients describing the degree of strictness to, or the room to manoeuvre from, these industry proportions are set, optimal shrinkage intensity parameters are chosen in the same way as before.

Shrinkage is equivalent to imposing penalty terms on the optimization objective. A practitioner's viewpoint would associate such penalty terms with the effect of transaction costs in portfolio selection. In fact, extending the problem to allow for transaction costs reduces the effect of the uncertainty regarding estimated optimization inputs on investment decisions. In a strange way, imposing a penalty improves the performance.

We follow a similar practice involving constraints on the absolute magnitude of the parameter vector in the linear model to express the portfolio selection problem with transaction costs. Once a link between the two problems has been established, an existing iterative optimization algorithm is modified to become suitable for portfolio selection. Provisions for a number of extensions, motivated by the asset allocation setting, are made. These include constraints on the individual asset weights, different base proportional costs to account for the effect of stock liquidity and increasing, possibly non-linear, marginal costs to reflect the effect of a trade made by a large institution which could result in a change in the share price. Finally, we show how this method could be used, under certain assumptions, with a different utility function.

Monte Carlo simulations and an analysis of a real-life dataset give support to our findings. Under different, simulated market scenarios, our proposed methods performed very well compared to existing procedures. They address the issue of uncertainty and consistently improve portfolio selection.

One possible shortfall in their performance is highlighted by the inclusion of transaction costs. Estimators which are not designed to take account of such costs are more likely to fail. Although this may be considered as a disadvantage, it also signals the direction of possible further research.

Appendix A

Portfolio Selection

A.1 Sharpe Ratio Optimization

We want to optimise:

$$g(\mathbf{w}) = \frac{(\mathbf{w}^T\boldsymbol{\mu})}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{1}{2}}}$$

By applying the chain rule we get:

$$\frac{\partial g}{\partial \mathbf{w}} = \frac{\boldsymbol{\mu}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{1}{2}}} - \frac{(\mathbf{w}^T\boldsymbol{\mu})\,\boldsymbol{\Sigma}\mathbf{w}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{3}{2}}}$$

$$\frac{\partial^2 g}{\partial \mathbf{w}\partial \mathbf{w}^T} = -\frac{\boldsymbol{\mu}\mathbf{w}^T\boldsymbol{\Sigma}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{3}{2}}} +$$

$$+\left[\frac{3\,(\mathbf{w}^T\boldsymbol{\mu})(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{1}{2}}\boldsymbol{\Sigma}\mathbf{w}\mathbf{w}^T\boldsymbol{\Sigma} - (\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{3}{2}}\left(\boldsymbol{\mu}\mathbf{w}^T\boldsymbol{\Sigma} + (\mathbf{w}^T\boldsymbol{\mu})\boldsymbol{\Sigma}\right)}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{3}}\right]$$

Hence, in general:

$$\frac{\partial^2 g}{\partial \mathbf{w}\partial \mathbf{w}^T} = -\frac{2\,\boldsymbol{\mu}\mathbf{w}^T\boldsymbol{\Sigma}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{3}{2}}} + \frac{3\,(\mathbf{w}^T\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{w}\mathbf{w}^T\boldsymbol{\Sigma}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{5}{2}}} - \frac{(\mathbf{w}^T\boldsymbol{\mu})\boldsymbol{\Sigma}}{(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})^{\frac{3}{2}}}$$

Now, let $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = x \lessgtr 0$ and $\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = y > 0$. The optimal $\mathbf{w}$ is the solution of $\frac{\partial g}{\partial \mathbf{w}} = \mathbf{0}$ and is given by:

$$\mathbf{w}^* = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{x}$$

Then $\mathbf{w}^{*T}\boldsymbol{\mu} = y/x$ and $\mathbf{w}^{*T}\boldsymbol{\Sigma}\mathbf{w}^* = y/x^2$. By substitution we get:

$$\left.\frac{\partial^2 g}{\partial \mathbf{w}\partial \mathbf{w}^T}\right|_{\mathbf{w}^*} = -2\left(\frac{y}{x^2}\right)^{-\frac{3}{2}}\frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{x} + 3\left(\frac{y}{x}\right)\left(\frac{y}{x^2}\right)^{-\frac{5}{2}}\frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{x^2} - \left(\frac{y}{x}\right)\left(\frac{y}{x^2}\right)^{-\frac{3}{2}}\boldsymbol{\Sigma}$$

$$= -2\frac{|x|^3}{x}y^{-\frac{3}{2}}\boldsymbol{\mu}\boldsymbol{\mu}^T + 3\frac{|x|^3}{x}y^{-\frac{3}{2}}\boldsymbol{\mu}\boldsymbol{\mu}^T - \frac{|x|^3}{x}y^{-\frac{1}{2}}\boldsymbol{\Sigma}$$
$$= -\frac{|x|^3}{x}y^{-\frac{1}{2}}\left(\boldsymbol{\Sigma} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{y}\right)$$

For a symmetric positive definite matrix $\mathbf{A}$ we have:

$$\left|\mathbf{A} + \mathbf{a}\mathbf{b}^T\right| = |\mathbf{A}|\left|\mathbf{I} + \mathbf{A}^{-1}\mathbf{a}\mathbf{b}^T\right| = |\mathbf{A}|\left(1 + \mathbf{b}^T\mathbf{A}^{-1}\mathbf{a}\right)$$

Using $\mathbf{A} = \boldsymbol{\Sigma}, \mathbf{a} = \boldsymbol{\mu}$ and $\mathbf{b} = -(1/y)\boldsymbol{\mu}$ results to:

$$\left|\boldsymbol{\Sigma} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{y}\right| = |\boldsymbol{\Sigma}|\left(1 - \frac{1}{y}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) = 0$$

Since $\left.\frac{\partial^2 g}{\partial\mathbf{w}\partial\mathbf{w}^T}\right|_{\mathbf{w}^*}$ is singular it is unclear whether the critical point $\mathbf{w}^*$ is a maximum, a minimum or a terrace point. However, we can inspect the diagonal elements:

$$\left(\frac{\partial^2 g}{\partial\mathbf{w}\partial\mathbf{w}^T}\right)_{ii} = -\frac{|x|^3}{x}y^{-\frac{1}{2}}\left(\boldsymbol{\Sigma} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{y}\right)_{ii} = -\frac{|x|^3}{x}y^{-\frac{3}{2}}\left(y\,\sigma_i^2 - \mu_i^2\right)$$

where μ_i^2 and σ_i^2 are the mean and variance of asset i for $i = 1, \ldots, N$. If *all* diagonal elements are negative then we can infer that the critical point $\mathbf{w}^*$ is a maximum of $g(\mathbf{w})$ whereas if *all* diagonal elements are positive then the critical point must be a minimum.

Assume initially that $\boldsymbol{\Sigma}$ is diagonal. Then

$$y = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \sum_{j=1}^{N}\frac{\mu_j^2}{\sigma_j^2}$$

and

$$\left(\frac{\partial^2 g}{\partial\mathbf{w}\partial\mathbf{w}^T}\right)_{ii} = -\frac{|x|^3}{x}y^{-\frac{3}{2}}\left[\sum_{j=1}^{N}\left(\frac{\mu_j^2}{\sigma_j^2}\right)\sigma_i^2 - \mu_i^2\right]$$
$$= -\frac{|x|^3}{x}y^{-\frac{3}{2}}\sum_{j\neq i}^{N}\left(\frac{\mu_j^2}{\sigma_j^2}\sigma_i^2\right)$$

Since $y > 0$, it is now clear that:

$$\left(\frac{\partial^2 g}{\partial\mathbf{w}\partial\mathbf{w}^T}\right)_{ii} < 0 \text{ if } x > 0$$

and

$$\left(\frac{\partial^2 g}{\partial\mathbf{w}\partial\mathbf{w}^T}\right)_{ii} > 0 \text{ if } x < 0$$

for $i = 1, \ldots, N$. Hence when $x = \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} > 0$, the Sharpe Ratio attains a *maximum* at $\mathbf{w}^*$ whereas when $x < 0$, $g(\mathbf{w})$ attains a *minimum*.

Since $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix, it can be diagonalized and written as $\boldsymbol{\Sigma} = \mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^T$. The proof remains the same for the now transformed data.

Appendix B

Taylor's Expansion

B.1 Second Order Approximation of $\hat{\mathbf{w}}_{SM}$

By replacing $\mathbf{S}$ with the population quantity $\mathbf{\Sigma}$, we approximate $\hat{\mathbf{w}}_{SM}$ by $\hat{\mathbf{w}}^*_{SM}$ where:

$$\hat{\mathbf{w}}^*_{SM} = \frac{\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}}.$$

Let $\mathbf{y} = \mathbf{\Sigma}^{-1}\bar{\mathbf{x}}$ and $z = \mathbf{1}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}$. Then, under the Normality assumption:

$$\mathbf{y} \sim \mathcal{MN}\left(\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \frac{1}{T}\mathbf{\Sigma}^{-1}\right)$$

and

$$z \sim \mathcal{N}\left(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \frac{1}{T}\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}\right)$$

We also have

$$Cov\left(\mathbf{y}, z\right) = \frac{1}{T}\mathbf{\Sigma}^{-1}\mathbf{1}$$

From the differentiable function $g(\mathbf{y}, z) = z^{-1}\mathbf{y}$ we can compute

$$\frac{\partial g}{\partial \mathbf{y}^T} = \frac{1}{z}\mathbf{I}, \quad \frac{\partial g}{\partial z} = -\frac{1}{z^2}\mathbf{y}$$

and

$$\frac{\partial^2 g}{\partial z^2} = \frac{2}{z^3}\mathbf{y}, \quad \frac{\partial^2 g}{\partial \mathbf{y}^T \partial z} = -\frac{1}{z^2}\mathbf{I}$$

Using the above, the second order multivariate Taylor expansion around the point

$$(\mathbf{y}_0, z_0) = \left(\mathbf{\Sigma}^{-1}\boldsymbol{\mu},\ \mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right)$$

is given by

$$\frac{1}{z}\mathbf{y} \approx \frac{1}{z_0}\mathbf{y}_0 + \frac{1}{z}(\mathbf{y}-\mathbf{y}_0) - \frac{z-z_0}{z_0^2}\mathbf{y}_0 + \frac{(z-z_0)^2}{z_0^3}\mathbf{y}_0 - \frac{z-z_0}{z_0^2}(\mathbf{y}-\mathbf{y}_0)$$

Since $E[\mathbf{y}] = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ and $E[z] = \mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$, the expectation is therefore given by

$$\begin{aligned}
E\left[\frac{1}{z}\mathbf{y}\right] &\approx \frac{\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}} + Var[z]\frac{\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^3} - \frac{1}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}Cov(\mathbf{y},z) \\
&= \frac{\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}} + \frac{1}{T}\left(\frac{\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}} - \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}\right) \times \frac{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}
\end{aligned}$$

The first order approximation for the variance is:

$$\begin{aligned}
Var\left[\frac{1}{z}\mathbf{y}\right] &\approx \frac{1}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}Var[\mathbf{y}] + \frac{Var[z]}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^4}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1} \\
&\quad - \frac{1}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^3}\left(Cov\left[\mathbf{y}, z\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right] + Cov\left[z\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \mathbf{y}\right]\right) \\
&= \frac{1}{T}\frac{1}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}\mathbf{\Sigma}^{-1} + \frac{1}{T}\frac{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^4}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1} \\
&\quad - \frac{1}{T}\frac{1}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^3}\left(\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\mathbf{1}^T\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-1}\mathbf{1}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\right) \\
&= \frac{1}{T}\mathbf{\Sigma}^{-1}(\mathbf{I}+\mathbf{K}) \times \frac{1}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}
\end{aligned}$$

with $\mathbf{I}$ being the identity matrix and

$$\mathbf{K} = \left(\frac{\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}}{(\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}\right)\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1} - \frac{\boldsymbol{\mu}\,\mathbf{1}^T\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-1}\mathbf{1}\,\boldsymbol{\mu}^T}{\mathbf{1}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}.$$

Appendix C

Median Results

Available in this appendix are the figures and tables corresponding to the ones presented in Chapter 3 using the affine equivariant median rather than the L_1 median. The results are almost identical.

C.1 Global Minimum Variance Portfolio

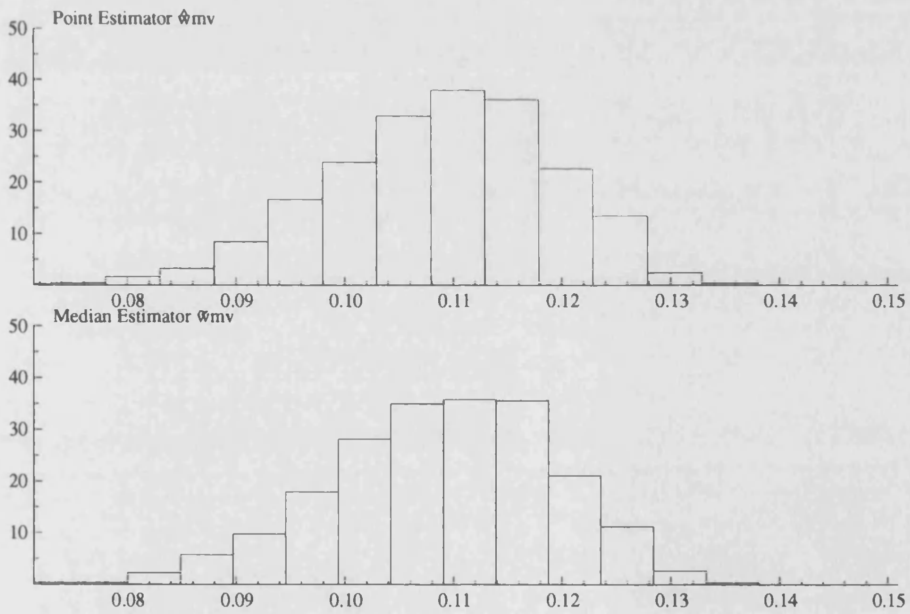

Figure C.1: Distribution of Sharpe Ratio of GMV Portfolio

	$SR(\hat{\mathbf{w}}_{GMV})$	$SR(\tilde{\mathbf{w}}_{GMV})$
Min	0.065	0.067
Q1	0.102	0.102
Mean	0.109	0.109
Median	0.109	0.109
Q2	0.116	0.116
Max	0.137	0.137
Res	478	522

Table C.1: Summary Statistics for Estimators of $SR(\mathbf{w}_{GMV})$

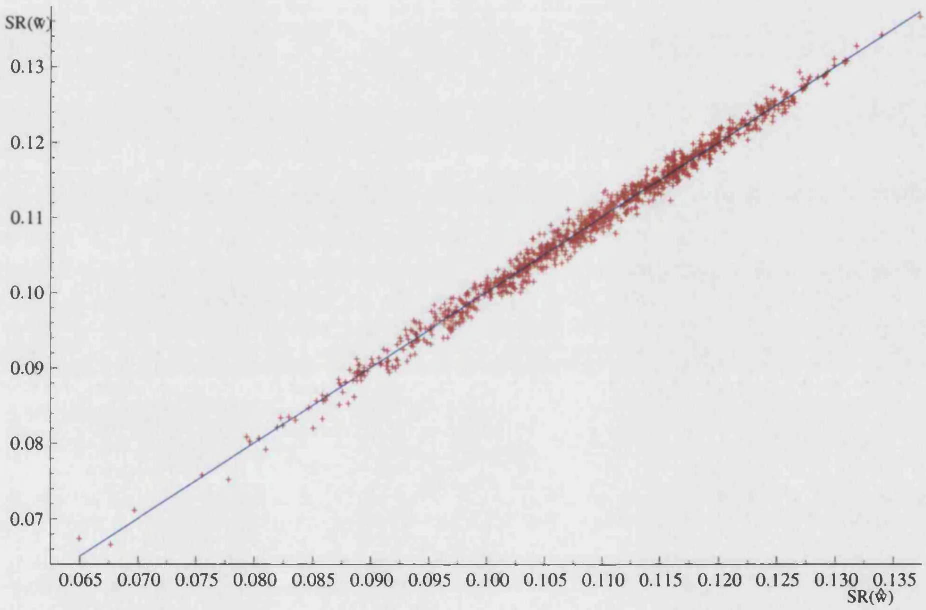

Figure C.2: Sharpe Ratio of GMV Portfolio under Two Estimators

C.2 Tangency Portfolio

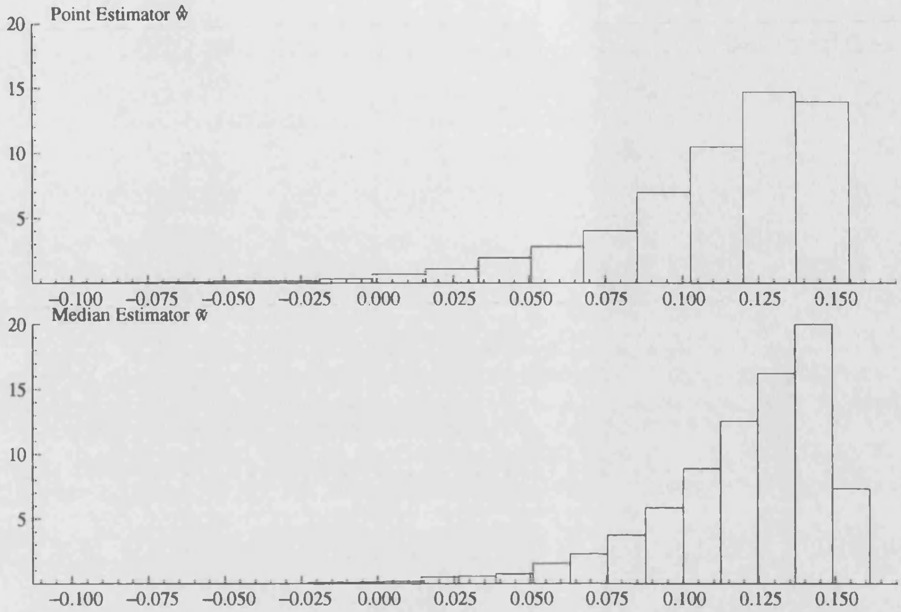

Figure C.3: Distribution of Sharpe Ratio of Tangency Portfolio

	$SR(\hat{\mathbf{w}}_T)$	$SR(\tilde{\mathbf{w}}_T)$
Min	-0.102	-0.016
Q1	0.092	0.108
Mean	0.110	0.121
Median	0.119	0.127
Q2	0.137	0.142
Max	0.155	0.155
Res	267	733

Table C.2: Summary Statistics for Estimators of $SR(\mathbf{w}_T)$

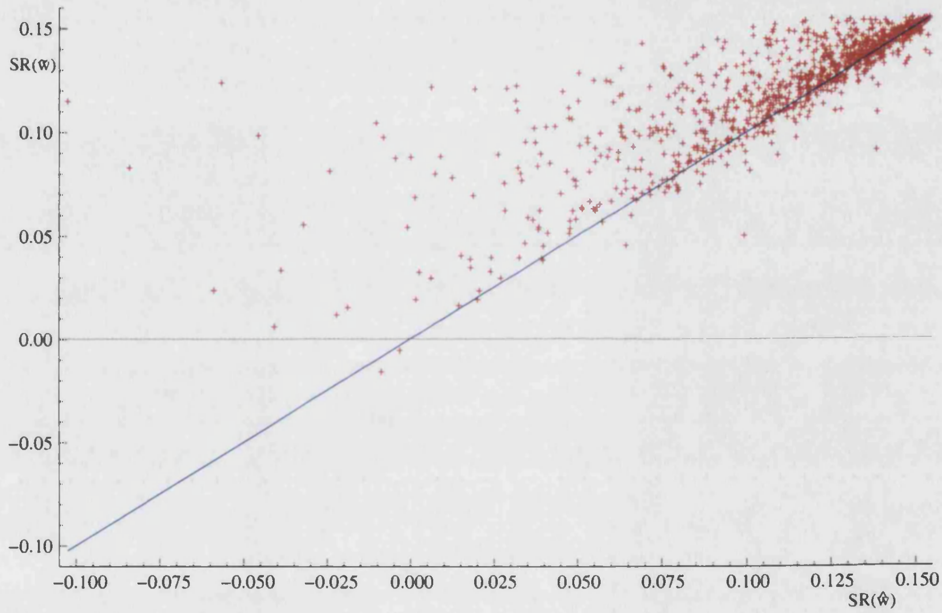

Figure C.4: Sharpe Ratio of Tangency Portfolio under Two Estimators

Appendix D

Programming

D.1 Ox Routines and GUI Application

A class of programming routines (`Markowitz` Class) was written in Ox 3.20 (Doornik, 2002) for the necessary calculations in this thesis. A Graphical User Interface (GUI) version is under development with OxJAPI[1], an Ox version of the Java Application Programming Interface. The source code for the programming routines is available from the author upon request.

The `Markowitz` Class can load either a sample of observations or the population moments of the assets' return distribution. It is easy to simulate a dataset from the population parameters, or display some summary statistics. In terms of plotting, the class includes methods for drawing the efficient frontier, utility curves (with either the negative exponential or the quadratic family) or specific portfolios on the portfolio return–standard deviation plane.

The class includes routines which calculate the estimators used in this research. The user has control over certain optional function arguments. For example, in order to find the lasso estimator, one can use built-in procedures to set the constraints or the transaction rate. Furthermore, some functions produce helpful graphical output.

Currently, the `Markowitz` Class can be included in Ox programs. Meth-

[1] Available from http://site.voila.fr/choirat/software/oxjapi/oxjapi.html.

ods from the class can be called from other programs. A more intuitive menu-driven User Interface is under development.

Bibliography

Bedall, F. K. and H. Zimmermann (1979). The Mediancentre. *Applied Statistics 28*(3), 325–328.

Bingham, N. H. and R. Kiesel (2002). Semi-Parametric Modelling in Finance: Theoretical Foundations. *Quantitative Finance 2*(4), 241–250.

Birge, J. R. and F. Louveaux (1997). *Introduction to Stochastic Programming.* New York: Springer-Verlag.

Breiman, L. (1996). Bagging Predictors. *Machine Learning 24*(2), 123–140.

Britten-Jones, M. (1999). The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights. *Journal of Finance 54*(2), 655–671.

Britten-Jones, M. (2000). Bayesian Portfolio Selection: A New Approach. Unpublished Report.

Brown, B. M. (1983). Statistical Uses of the Spatial Median. *Journal of the Royal Statistical Society, Series B 45*(1), 25–30.

Bühlmann, P. and B. Yu (2001). Analyzing Bagging. *Annals of Statistics 30*(4), 927–961.

Chakraborty, B., P. Chaudhuri, and H. Oja (1998). Operating Transformation Retransformation on Spatial Median and Angle Test. *Statistica Sinica 8*, 767–784.

Chaudhuri, P. (1992). Multivariate Location Estimation using Extension of R-Estimates through U-Statistics Type Approach. *Annals of Statistics 20*(2), 897–916.

Cox, D. R. (1975). Partial Likelihood. *Biometrika 62*(2), 269–276.

Doornik, J. A. (2002). *Object-Oriented Matrix Programming Using Ox* (3rd ed.). London: Timberlake Consultants Press. Also Oxford: www.nuff.ox.ac.uk/Users/Doornik.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics 7*(1), 1–26.

Efron, B. and C. Morris (1976). Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Annals of Statistics 4*(1), 11–21.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Fang, K. T., S. Kotz, and K. W. Ng (1990). *Symmetric Multivariate and Related Distributions.* New York: Chapman and Hall.

Freidrich, C. J. (1929). *Alfred Weber's Theory of Location of Industries.* University of Chicago Press.

Friedman, J. H. and P. Hall (2000). On Bagging and Nonlinear Estimation. Unpublished Manuscript. Department of Statistics, Stanford University.

Frost, P. A. and J. E. Savarino (1986). An Empirical Bayes Approach to Efficient Portfolio Selection. *Journal of Financial and Quantitative Analysis 21*(3), 293–306.

Fu, W. J. (1998). Penalized Regressions: The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics 7*(3), 397–416.

Gower, J. C. (1974). The Mediancentre. *Applied Statistics 23*(3), 466–470.

Hettmansperger, T. P. and R. H. Randles (2002). A Practical Affine Equivariant Multivariate Median. *Biometrika 89*(4), 851–860.

Ingersoll, J. E. (1987). *Theory of Financial Decision Making.* Maryland: Rowman & Littlefield.

Jagannathan, R. and T. Ma (2002). Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. Working Paper 8922, NBER.

James, D. and C. Stein (1961). Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* Volume 1, Berkeley, pp. 361–379. University of California Press.

Jobson, J. D. and B. Korkie (1983). Statistical Inference in Two-Parameter Portfolio Theory with Multiple Regression Software. *Journal of Financial and Quantitative Analysis 18*(2), 189–197.

Jorion, P. (1986). Bayes-Stein Estimation for Portfolio Analysis. *Journal of Financial and Quantitative Analysis 21*(3), 279–292.

Kall, P. and S. W. Wallace (1999). *Stochastic Programming.* New York: John Wiley.

Ledoit, O. (1995). A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. Working Paper 24-95, The Anderson School, UCLA.

Ledoit, O. (1997). Improved Estimation of the Covariance Matrix of Stock Returns With an Application to Portfolio Selection. Working Paper 6-97, The Anderson School, UCLA.

Liu, R. Y., J. M. Parelius, and K. Singh (1999). Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference. *Annals of Statistics 27*(3), 783–840.

Lopuhaä, H. P. and P. J. Rousseeuw (1991). Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *Annals of Statistics 19*(1), 229–248.

Luenberger, D. G. (1997). *Investment Science.* Oxford University Press.

Markowitz, H. M. (1952). Portfolio Selection. *Journal of Finance 7*(1), 77–91.

Markowitz, H. M. (1991). *Portfolio Selection: Efficient Diversification of Investment.* Cambridge, Massachusetts: Blackwell.

Michaud, R. O. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation.* Boston: Harvard Business School Press.

Milasevic, P. and G. R. Ducharme (1987). Uniqueness of the Spatial Median. *Annals of Statistics 15*(3), 1332–1333.

Rubin, D. B. (1981). The Bayesian Bootstrap. *Annals of Statistics 9*(1), 130–134.

Sharpe, W. (1963). A Simplified Model for Portfolio Analysis. *Management Science 9*(2), 277–293.

Sharpe, W. F. (1966). Mutual Fund Performance. *Journal of Business 39*(1), 119–138.

Small, C. G. (1990). A Survey of Multidimensional Medians. *International Statistical Review 58*(3), 263–277.

Stein, C. (1956). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* Volume 1, Berkeley, pp. 197–206. University of California Press.